CONCIERTO | SOVEREIGN INTELLIGENCE

Sovereign Perspectives · Sustainable and Resilient AI

Efficiency Versus Consumption: It Is in the Design

Why a Metered Model Bills for Inefficiency, and an Owned One Is Built to Avoid It

Series Sovereign Perspectives
Category Sustainable & Resilient AI

Format Analysis
Published by Concierto



Version 1.0 · September 2026

Copyright © 2026 Concierto. All rights reserved.

Contents

Executive Summary

1. The Wall at the End of Scale
2. The Consumption Paradox
3. Efficiency Is a Design Choice
4. If This Goes Unattended
5. The Path Forward: Efficiency by Design
6. What Leaders Must Do
7. The Reframe

About This Series · Sources

Executive Summary

The industry sells its models as always improving. The consumption bills tell the opposite story: whatever the user experiences, the meter only ever climbs. That contradiction is not an accident. It is the design.

The prevailing model of AI is built on brute-force scale, and it is meeting the limits of that approach on three fronts at once: diminishing returns on capability, and rising walls of energy, carbon, and cost. Providers answer with a story of continual, experience-based improvement, yet the consumption patterns underneath tell a different and more revealing story: usage, and the bill, rise regardless of how good the experience is said to be, because a consumption-metered model profits when computation increases, not when it decreases. A sovereign, purpose-built model inverts the incentive. It is designed to reach the most accurate answer as fast as possible, so it uses less, and efficiency by design turns out to serve cost and the environment at the same time.

280x	fall in inference cost, two years
40%	annual gain in model efficiency
142x	smaller model matching a giant

In Brief

- Brute-force scaling is hitting diminishing returns and energy, carbon, and cost walls simultaneously.
- Public models advertise learning-based improvement, but consumption and the bill rise no matter what the user experiences.
- A consumption meter rewards more computation, so inefficiency is not a bug in that model; it is the revenue.
- A sovereign, purpose-built model is designed for the fastest accurate answer, so sessions are shorter and infrastructure load falls.
- Token efficiency by design serves cost sustainability and environmental sustainability together.

A model paid by the token improves your experience and its own revenue at the same time, by making you consume more. The only model that is paid to make you consume less is one you own.

Copyright © 2026 Concierto. All rights reserved.

1. The Wall at the End of Scale

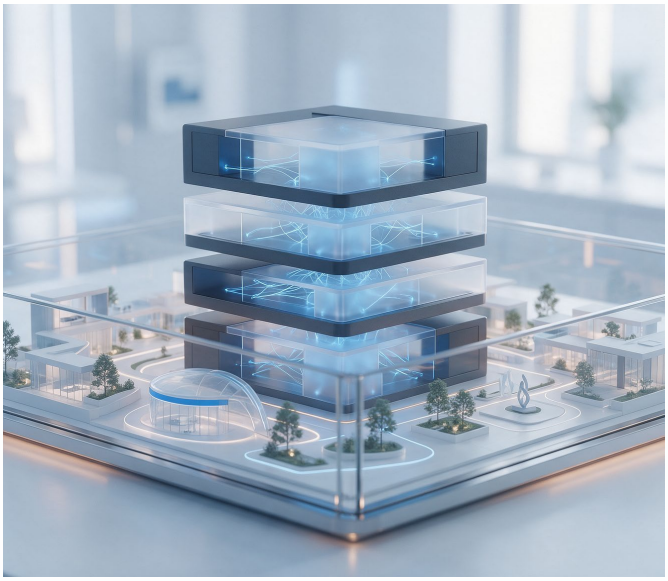
The brute-force paradigm assumed that more, more parameters, more compute, more tokens, would keep buying more capability. It no longer does.

The evidence that scale has entered diminishing returns is now firm. Capability from raw size follows a flattening curve, and the practical proof is that far smaller models have caught the giants on real tasks, one recent model matching the performance of a system roughly a hundred and forty times its size. At the same time the cost of a given level of capability has collapsed, inference cost falling by orders of magnitude in two years, and efficiency has improved by roughly forty percent a year. The frontier of value has moved off raw scale and onto efficiency and fit.

142x

A model 142 times smaller now matches the performance of a leading giant, which means the advantage has left raw scale behind.





2. The Consumption Paradox

Here is the contradiction the industry does not advertise: the models are sold as improving, and the bills only rise.

Providers describe their systems as continually learning and improving with use, and users often experience exactly that. But the consumption data underneath tells a story the improvement narrative obscures: usage climbs, and the bill climbs with it, no matter what the experience is said to be. This is not a paradox once the incentive is named. A consumption-metered model earns more when more computation happens, so every part of the system, from verbose answers to needless re-tries to open-ended exploration, works in the provider's favour. Improvement and consumption are sold as the same thing, when in the metered model they are simply both going up, and only one of them is going up for the customer's benefit.

Copyright © 2026 Concierto. All rights reserved.

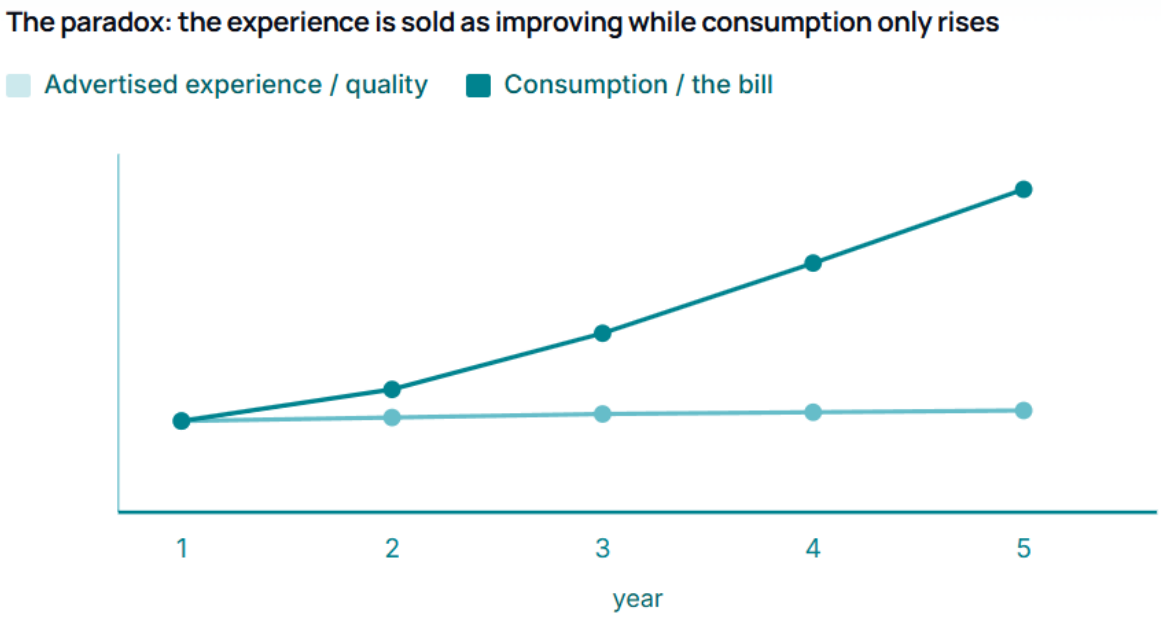


Figure 1. The experience is sold as improving while consumption only rises. Illustrative.

The meter
A model paid by consumption is paid for inefficiency, the one arrangement in a market economy that rewards using more to do the same.

3. Efficiency Is a Design Choice

Efficiency is not something a model drifts toward. It is something a model is either designed for or designed against.

A sovereign, purpose-built model is designed to do the opposite of what a consumption meter rewards: to reach the most accurate answer as quickly as possible. Because it is fitted to a domain rather than built to be universal, it carries less noise, needs less prompting to be corrected, and reaches the first-time-right answer more often. The consequence shows up directly in the numbers that matter operationally: fewer computations per resolved task, shorter sessions, and lower load on the infrastructure underneath. Token efficiency, in other words, is not a tuning parameter; it is a property of designing the model for the work rather than for the meter.

Copyright © 2026 Concierto. All rights reserved.

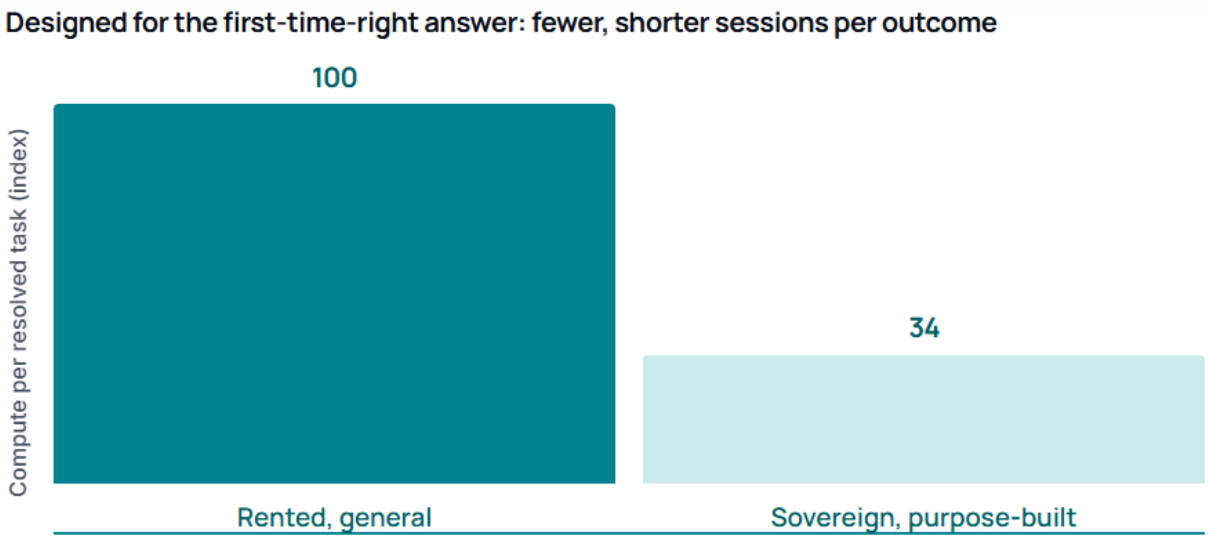


Figure 2. Designed for the first-time-right answer: fewer, shorter sessions per outcome. Illustrative.

40%/yr
Efficiency is improving about forty percent a year, so a model designed to capture that curve pulls away from one designed to bill against it.

4. If This Goes Unattended

Leave the consumption model in place and the costs do not plateau as the technology matures. They compound, in money and in carbon.

If the metered model remains the default, three costs rise together and without limit. The financial cost climbs because the incentive is to consume more, and agents, which run continuously, multiply that consumption far beyond what human use ever did. The energy and carbon cost climbs with it, because every unnecessary computation is also unnecessary power drawn and emissions produced, and the brute-force build-out assumes ever more of both. And the strategic cost climbs, because an organization whose AI spend rises with use cannot forecast it, cannot cap it without rationing the capability, and cannot escape it without leaving the model it built on. Unattended, the consumption model turns a maturing technology into a growing liability.

Copyright © 2026 Concierto. All rights reserved.

In every other technology, maturity brings falling unit costs. The consumption model is engineered so that maturity brings rising bills, because the bill is the product.

5. The Path Forward: Efficiency by Design

The way out is not to use the metered model more carefully. It is to run a model whose design goal is to consume less.

An owned, sovereign, purpose-built model aligns the incentive with the customer for the first time. Because it is not billed by consumption, there is no reward for inefficiency and every reason to design for the fastest accurate answer. Because it is purpose-built, it reaches that answer with less computation. And because it is efficient, it lowers cost and footprint in the same motion, since a computation not performed is both a dollar not spent and a watt not drawn. The environmental case and the financial case, so often treated as a trade-off, turn out to be the same case: fewer, better, first-time-right computations are cheaper and cleaner at once.

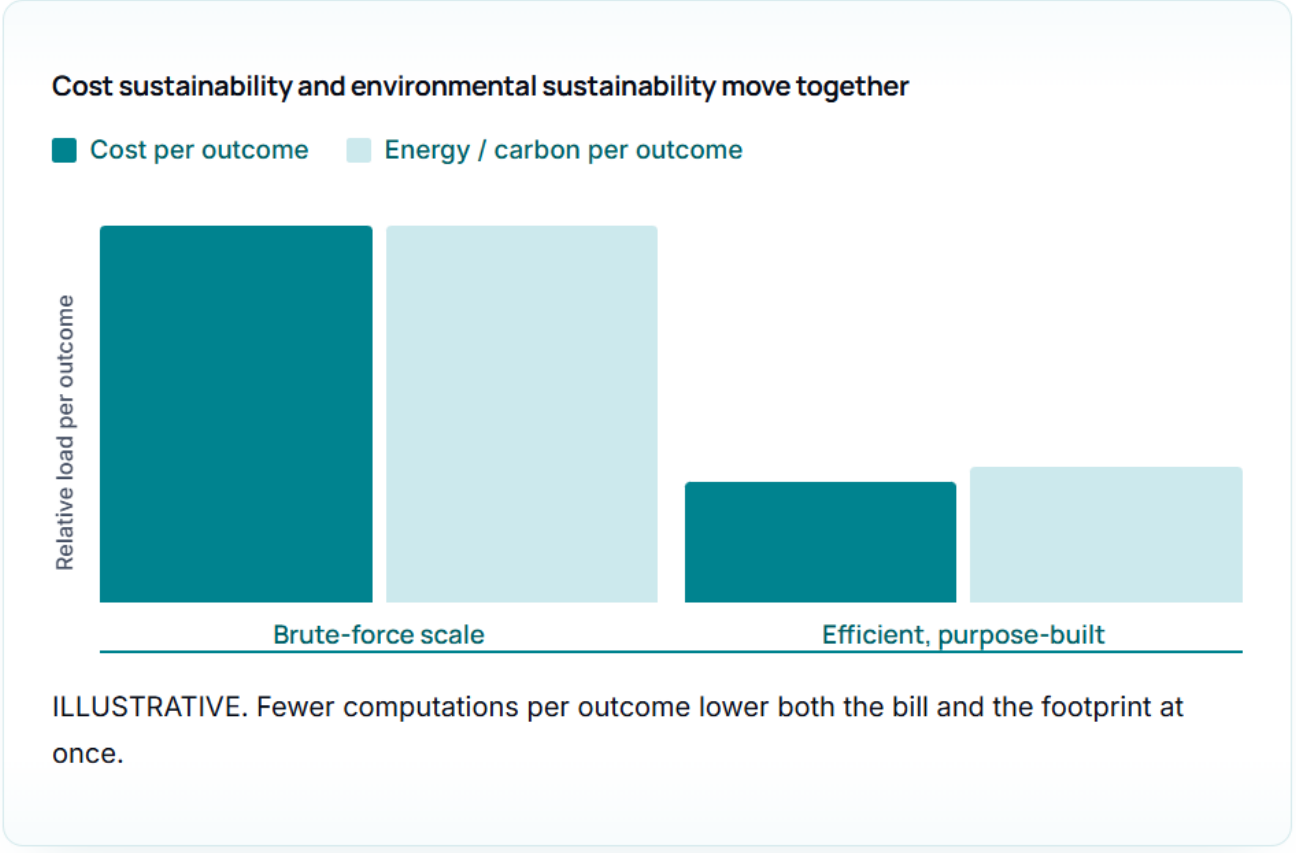


Figure 3. Cost sustainability and environmental sustainability move together when the model is efficient by design. Illustrative.

Copyright © 2026 Concierto. All rights reserved.

The greenest computation and the cheapest computation are the same one: the one you did not need to perform because the answer was right the first time.

Reject the consumption incentive for anything at scale, because a model that profits from your inefficiency will never be optimized for your economy. Favour purpose-built, owned models that are designed for the first-time-right answer, and treat the resulting reduction in cost and carbon as a single, linked benefit rather than two separate initiatives.

6. What Leaders Must Do

The correction is to change what is measured, from raw scale and capability to efficiency and outcome per unit of compute.

Measure efficiency, not size: the meaningful metric is outcome per unit of compute, not parameters or peak benchmark scores,

280x

Inference cost has fallen 280-fold in two years; an efficient, owned model captures that gain, while a metered one quietly keeps the savings.

Action	What it delivers
Measure outcome per unit of compute	The metric rewards efficiency, not raw scale
Reject the consumption incentive at scale	The provider is no longer paid for your inefficiency
Choose purpose-built, owned models	Fewer computations per resolved task; shorter sessions
Treat cost and carbon as one benefit	Efficiency lowers both in the same motion
Design for first-time-right	The cheapest and cleanest computation is the one avoided

Copyright © 2026 Concierto. All rights reserved.

7. The Reframe

The question was never how much intelligence you can afford to consume. It is how little consumption a given outcome requires.

The consumption model asks the customer to buy ever more computation and calls the rising bill a sign of engagement. The efficiency model asks how few computations an outcome truly needs and treats every one avoided as a win for the budget and the grid alike. Only one of these is designed in the customer's interest, and it is the one the customer owns. Efficiency is not a feature to switch on. It is in the design, and the design is either built for your economy or for the meter.

Stop asking how much AI you can consume. Start asking how little a result requires. The model that answers the second question is the only one on your side.

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, Artificial Intelligence Index Report 2025. hai.stanford.edu.
- OECD.AI Policy Observatory, compute, efficiency, and environmental indicators, 2024–2026. oecd.ai.
- International Energy Agency analysis of data-centre and AI electricity demand, 2024–2026. iea.org.

reach@concierto.cloud

www.concierto.cloud

www.trianz.com

Copyright © 2026 Concierto. All rights reserved.