Scarcity Counts the Order, Not the Consumption

When capital, architecture, and value decouple, resource allocation follows whichever market still emits a price. Enterprise AI is the current instance.

The deployability thesis held that the hard problem of enterprise AI is not the model but its institutional insertion.

This piece states the mechanism the earlier work circled, at the level where it is reusable rather than the level where it is dramatic. When three markets that should align come apart, a capital market that prices future capacity, an architecture market that prices what gets built, and a value market that would price realized outcomes if it could, resource allocation is governed by whichever market still emits a price signal.

In enterprise AI in mid-2026 that signal comes disproportionately from the capital market, because the value market has no standardized instrument of measurement yet. The GPU, the token, the agent, the scarcity, the customer ROI, are the mechanism and its symptoms. The claim, stated generally, is that this happens wherever value creation is deferred or hard to measure and capital is not, which is why it should be readable outside AI.

Two terms have to be pinned before they drift.

- First, by *rent* this article means the operating cost incurred above what the task requires to reach the same useful decision. The surplus, and only the surplus, not the agent, not the complexity a problem genuinely demands,
- Second, *disproportionately* is deliberate. The architecture market answers to several signals at once, talent, regulation, open standards, cloud economics and the price of compute. Capital is simply the loudest of them today, not the only one.

The title's claim is a consequence, and a compression. Scarcity is overdetermined, reflecting anticipations, real demand, build plans, industrial constraints and expected future consumption at once; the formula isolates the single component the bullish reading treats as the whole, the order placed against a plan, and insists only that it not be mistaken for the served workload. The domain of validity is narrow: enterprise deployment under usage-based billing, on the realized figures of mid-2026, strongest where the purchase is a committee decision and the signer is insulated from the meter, weakest where a single owner prices its own cash flow. It is not a bubble call; at least one

frontier lab projects an operating profit in the second quarter of 2026 on gross margins near seventy-seven percent.

I. Capital allocation: who sets the half-life of the asset

Hyperscalers spend at a rate no near-term revenue justifies because the market rewards positioning over profitability. The five largest are on course for roughly seven hundred and twenty-five billion dollars of capital expenditure in 2026, part of a deployment Allianz Research sizes near two point one trillion across 2026 to 2028, at a capex intensity of thirty-four percent of revenue, double the fifteen percent peak of the late-1990s internet build, and held there regardless, which is the tell. When the market punishes whoever blinks first, building ahead of demand is defensive discipline, and no one need be scheming to dump compute for the overbuild to occur.

Some actor sets the clock that turns overbuild into urgency, and today it is mostly Nvidia. Intel sold processors; Nvidia currently sets the economic half-life of the data center, each architecture it ships, Hopper then Blackwell then the Rubin generation shown at GTC 2026, compressing the economic life of the one before it. The position is contested rather than owned: AMD, Broadcom, Google's TPU, Amazon's Trainium and the inference-specialized silicon of Groq and Cerebras all press on the pace it sets. The structural point survives the competition. Whoever sets the depreciation clock forces the owner to fill capacity before it expires, and the round-tripping the discourse ascribes loosely to the sector lives here, in the capital flows where an investor's cash returns as its own revenue, far more than in the labs' operating revenue.

II. Infrastructure economics: what the shortage measures

Two facts sit in the same market. High-bandwidth memory and packaging are effectively sold out, GPU lead times run in quarters, and power has replaced silicon as the binding constraint, with the Apollo and JPMorgan desks describing a supply chain tight through 2026. Against that, Cast AI's 2026 review of roughly twenty-three thousand enterprise Kubernetes clusters found average GPU utilization near five percent.

The five percent must be handled with its leash on. It measures *enterprise provisioned Kubernetes capacity*, not global compute, and not the labs' production inference, which runs hot; a Constellation Research analyst rightly cautioned that the sample skews toward test and poorly optimized estates. Read at its scope it proves one thing: a large population of enterprise clusters, provisioned during the scramble, sits idle while the provider bills for it whether it turns or not.

The apparent paradox dissolves once one asks what the scarcity is measured against. Memory is sold out against orders, and orders are placed against build targets. The strongest counter deserves full weight: fiber in the late 1990s and cloud in the late 2000s

ran at abysmal initial utilization and preceded their applications by years, so overcapacity may be the normal lag of a transition whose demand is latent. The disanalogy is the depreciation clock of the previous layer. Fiber was cheap to hold idle; GPU capacity is not, written down in eighteen to thirty-six months. The lag reading and the thesis diverge on how long the gap can persist before it is written off, which is why the gap is the number to watch: Sequoia's David Cahn puts the annual distance between spend and ecosystem revenue near six hundred billion dollars and widening.

The buyer's price intuition deserves a more cautious note than a cynic would write. It is tempting to say the enterprise is fooled by the sponsored consumer product, the free tier where roughly ninety-five percent of one assistant's users pay nothing, into thinking intelligence is cheap. That overstates it. The 2026 buyer is often the same person who has read an AWS bill for anomalies for a decade. The gap that matters is not price literacy, which the mature buyer has, but the absence of a quality metric, which no one has: a CIO can read the token bill to the cent and still hold no instrument that tells whether the architecture generating it was sized to the problem.

III. Software architecture: the five sources of weight

A chatbot query is one inference; an agentic workflow that plans, calls tools, verifies and self-corrects fires ten to twenty calls to finish one task, a real change of scale and not the thousandfold figure that circulates without a source. Multi-agent orchestration is not, in 2026, a billing artifice. It is the working answer to the reasoning limits of a monolithic model. The distinction is not *agent* against *no agent*. It is *consumption* against *productive consumption*.

Weight has five sources, and only one is the target. It can be a property of the problem: identity, security, observability and orchestration are hard, and a serious deployment is heavier than a demonstration. It can be a property of the *regulation*: under the AI Act or the European Health Data Space, an architecture that logs, isolates, documents and keeps a human in the loop is heavier than one that does not, and that weight is legal optimality, rational on an axis that is not the economic one. It can be *optionality*, and of a particular kind the finance view misses: an oversized deployment may be rational not for the current use case's return but as an entry ticket, the price of stabilizing internal skills, restructuring the enterprise's data and working the kinks out of a capability the organization will need next. That surplus is tuition, not rent, and only the buyer knows which it bought. It can be *immaturity*, an agent stacked on an agent by a team under deadline. And it can be *rent*, the surplus as defined. Four of the five are legitimate. The argument concerns the fifth, and its entire difficulty is that the fifth is indistinguishable from the other four without a baseline almost no one runs, the cheaper architecture set alongside to see whether it would have sufficed. The thesis does not claim enterprise

architectures are too heavy. It claims that nothing in the buyer's instruments separates the weight that is warranted from the weight that is rent.

A single illustration fixes the point without carrying the argument. A calibrated toxicity model mapping a molecule to fourteen endpoints, of the kind built as ToxTwin on the Institute's own terrain, is a computation that predicts, reports a confidence and stops, at a fraction of what a generic agent would spend to reach the same output. It is one case where the baseline is obvious. In most enterprise settings it is not, and that, not the existence of the frugal option, is the problem.

IV. Enterprise deployment: two incentives, one direction

The architecture reaches the client through the *deployer*, which is not one actor but two utility functions that happen to align. The hyperscaler's professional-services arm inherits the owner's imperative, utilization. The system integrator, Accenture, Capgemini, Deloitte and the rest, runs on a different function, billable hours, maintenance, support and change management, none of which reward the lighter architecture, and it is often the integrator, closer to the client's estate and paid by the complexity of the integration, who is the more decisive of the two. Neither requires bad faith. Even where the complexity is entirely legitimate, neither faces an economic incentive to seek simplicity, and the bias survives everyone's good intentions because it is a property of the incentive field, not of anyone's character.

Set the deployed narrative against the record. A deployment is sold as a productivity gain, a figure of forty percent on some team. MIT's NANDA initiative, in *The GenAI Divide: State of AI in Business 2025*, found about ninety-five percent of generative-AI pilots with no measurable impact on the profit-and-loss statement and about five percent reaching rapid revenue acceleration; Boston Consulting Group's *Widening AI Value Gap* of September 2025 put sixty percent of firms at no material value; McKinsey's late-2025 survey found enterprise EBIT impact at fewer than four in ten adopters. The narrative survives the record because the cost it omits is institutional: evaluation, observability, governance and the human review a confidently wrong system forces. An architecture can be cheap in tokens and ruinous in organization, and a customer-ROI figure that counts neither is not measuring a return.

V. The missing ruler, and why it resists being built

None of the above closes a deal if the buyer holds an instrument that prices weight against value. Today it mostly does not. George Akerlof, in "The Market for Lemons" (1970), showed that when buyers cannot observe quality the market selects for what is observable, and unseen quality cannot be paid for; here the unobservable is the line between productive consumption and rent, the observable is the demonstration and the

benchmark, and the market selects for what the supply side is rewarded to show, which is weight.

The obvious rejoinder is that if the ruler were so valuable the market would have built it, so its absence must be temporary or illusory. The truth is that the ruler is genuinely hard to build, for four reasons that compound. Causality is diffuse: the value of a good decision spreads across teams and quarters and rarely traces to a line. Temporality is long: the return on restructuring a workflow surfaces years after the compute bill. Attribution is near-impossible: when a human validates the model's output, the value of the decision cannot be cleanly split between the model and the expert who corrected it. And the externalities are organizational: the gains land on morale, retention, speed and optionality, none of which finance books. A market does not fail to build an instrument out of laziness. It fails because the quantity resists measurement, and the ruler has to be forged against that resistance rather than adopted off a shelf.

It is being forged all the same, which is why the failure is a phase and not a fate. FinOps arose to discipline cloud spend, LLMOps and model observability to instrument inference, benchmarking and process mining to attach numbers to outcomes, and on the supply side open-weight and small models, distillation and on-premises inference let an enterprise escape the metered pipe entirely. The claim is therefore dated: the pressure is real now, while the instrument is immature, and it recedes as the instrument standardizes. The historical parallel needs the same restraint. Enterprise resource planning ran on lock-in, cloud on economies of scale, customer relationship management on weak network effects, enterprise AI on variable billing; these are not the same market, and what they share is only that at purchase the economic quality could not be measured, so each was bought on faith before it was rationalized later, the condition Solow named in 1987 as the computer age visible everywhere except in the productivity statistics. One thing is new: under a license the unmeasured period cost was fixed, and under usage-based billing it is not. A wave you can rationalize later is a wave whose meter does not run while you wait. This one runs.

The instrument, once named, is not one instrument but a family, and not a neutral one. Call it *cost per useful decision*: a ratio whose denominator counts decisions whose downstream value exceeds the cost of verifying them, whose numerator sums the total lifetime cost of producing them, compute plus human oversight plus governance plus verification plus operations, carrying an option term for the future adaptation a heavier architecture genuinely buys. The denominator is domain-specific and cannot be otherwise, a useful decision being correct and defensible for a legal assistant, timely and safe for a triage system, accurate enough to act on for a forecast. And here is its sharpest limit: whoever defines the useful decision controls the ratio, so if the integrator captures that definition the metric becomes marketing rather than discipline, and the asymmetry

simply moves from the token to the word *useful*. The ruler is real and worth building. The contest over its denominator is the next place the rent will try to hide.

VI. The contribution, its reach, and its proof

The deepest version of the claim is a decoupling of technical performance from economic performance. There was a time when technical performance translated into economic value directly enough that optimizing the first served the second; that translation has thinned, and between the two the market now interposes consumption, then expected value, then a hypothetical return, optimizing the first link because it is the only one it can read. The three markets are tangled where they should be distinct, and the value market emits no price, so the architecture market is disproportionately driven by the capital market instead. This is not peculiar to AI. Wherever a capital market prices future capacity, an architecture market prices what gets built, and a value market cannot yet price outcomes, the second is governed increasingly by the first, a pattern visible in outline across biotechnology platforms, energy infrastructure, quantum and defense procurement. AI is the instance where the meter runs fastest, which is why it shows the pattern first and worst.

One tension has to be resolved rather than smoothed, because two failures are in play and only one is transitory. The buyer's inability to measure is transitory: it closes as the ruler is built. The vendor's inability to guarantee a probabilistic outcome may be permanent: a non-deterministic system resists the contractual guarantee that outcome-based pricing requires, so token pricing may not vanish at maturity, and the buyer may internalize architecture risk for good. These are different things. The market failure this article describes, the buyer signing without a ruler, is a phase. The pricing structure that leaves architecture risk with the buyer may be a permanent feature of a probabilistic technology. A mature market is therefore not one where the vendor bears the risk, but one where the buyer bears it knowingly, with an instrument to price it.

A word on what this argument is. It proposes a mechanism and illustrates it; it does not measure it. Akerlof, Solow, Cast AI, MIT, Sequoia and the rest illustrate a chain that has not been tested as a chain. The independent variable is stated plainly, does the buyer possess a quality metric, and the dependent variables follow, architectural complexity, operating cost, realized return; the claim is that the first governs the rest while the metric is absent, and the honest status of that claim is a hypothesis with a specified test, not a result. The next step is not another argument but a protocol that measures cost per useful decision across matched deployments with and without the metric and reads the difference. Until then this is a theory with good illustrations, a smaller thing than a proof and a larger thing than an essay.

For the decision-maker the consequence is neither adoption on faith nor withdrawal in panic, the same error facing opposite directions. It is to build the ruler faster than the

market standardizes it, and to refuse the two substitutions the current phase runs on: the shortage against a plan read as a shortage against a need, and the committed spend read as a realized return. The organizations that priced their deployments in useful decisions will still be standing when the meter has run its course. The organizations that priced them in conviction will discover what the invoice was for.

A customer ROI that is asserted and never measured is not a return. It is a narrative with an invoice attached.