



STRATEGIC ESSAY

Editorial date: 24 October 2026

The Computer That Cuts Its Own Wage

Why might the price of intelligence fall before the machine has paid for itself?

The Token Deflation Trap: When Faster Inference Undermines Compute Payback

Opening: A processor can be excellent at the job it was bought to perform and economically behind on the very same day. This is not a puzzle about faulty chips. It is about an industry capable of increasing its output of intelligence faster than its ability to charge for it. We open an American financial filing, move to an algorithm modeled on memory pages, and then to computing infrastructure that has entered service in Saudi Arabia. What emerges is that the decisive contest may concern the moment when computation becomes useful and saleable, rather than the number of calculations theoretically possible.

1. On the accountant's books, a chip lives for six years; in the market, a new rival appears every quarter

CoreWeave's 2025 annual report contains a small sentence a reader could pass over like an employee finding no surprise in a payslip: from January 2023, the company changed the estimated useful life of its data-center computing equipment from five years to six. Its disclosure attributes the change to advances in hardware, software improvements, and data-center design. This is an **estimated accounting life** used to allocate depreciation. It is not a promise that the equipment will remain the most profitable hardware for six years S01.

On 5 January 2026, NVIDIA introduced its Rubin platform and said its integrated design could reduce **inference token generation cost** by as much as tenfold relative to Blackwell for certain workloads. This is a vendor announcement describing a targeted technical capability. It is neither a disclosure of actual selling prices nor a judgment covering every model and application S03.

Then came the first-half 2026 filing. CoreWeave said that its remaining performance obligations (RPO) at the end of June stood at **\$103.7 billion**. The number is striking, but its definition matters more: it is contractual value allocated to services not yet delivered, subject to accounting rules and possible effects from delays and concessions. It is neither cash in the bank nor guaranteed net profit. The company expected to recognize about 41% of it over the following twenty-four months, then 39% over the twenty-four months after that S02.

Discussions of AI computing usually begin with the chip's name and memory capacity. Yet a file containing "six years," "a tenfold reduction in token cost," and "\$103.7 billion in future services" presents a more intriguing and more dangerous question: **which of these clocks must win the race?** The depreciation clock? The arrival of a new technology generation? The duration of contracts protecting cash flow? Or the pace at which the price a customer will pay changes?

None of these figures proves a market collapse or shows that any particular company is caught in a trap. Long-term contracts are, on the contrary, an important reason to resist the story that a processor becomes worthless the moment its successor arrives. The figures do show that the right investment question is not "Does the chip work?" It is "Who will buy its operating hours, at what price, how often, and before which capital obligation comes due?"

2. News that reaches a developer as a discount may reach a financier as a new assumption

In September 2026, OpenAI announced that the application programming interface (API) prices for GPT-6 Sol and GPT-6 Luna were lower than the corresponding promotional prices of the preceding generation, attributing the change in part to improvements in caching and inference. In the published example for Sol input, the price fell from **\$4 to \$2 per million tokens**, and output fell from **\$20 to \$10**. This compares two products from different generations under published pricing terms. It does not prove that the market rental rate of every chip was cut in half S04.

Someone running a small application may see room to experiment where inference bills once imposed a limit: testing ten formulations instead of five, or serving a user whose conversation could not previously justify a more capable model. Someone who bought racks of processors

and turned their price into a long financing obligation will ask a different question: will cheaper service cause demand to double just as quickly, or will more work have to be sold to earn the same revenue?

This is where what I call the **Token Deflation Trap** begins. A lower price alone does not spring the trap. Three things must coincide: revenue per unit of work sold falls; the number of useful, paid units cannot rise far enough; and capital comes due without its cost falling at the same pace. If any one condition fails, the trap may become a remarkable opportunity. The name is therefore useful as a question aimed at the critical joints, not as a prophecy.

The market is also more complicated than “price per million tokens.” An input token is not an output token; a cached input token whose context is reused is not fresh computation; and a successful concise answer may eliminate a long chain of attempts. If we sell a million tokens as though we were selling a million identical nails, we will discover that some nails consume memory, others require waiting time, and others came out of the cache in the first place. The accounting unit is convenient. The physical unit refuses to behave.

3. The token’s journey, hidden from the chat screen

To understand how software can reprice hardware, begin with an ordinary user’s sentence: “Summarize this correspondence and connect the financial decision to last week’s messages.” A graphics processing unit (GPU) does not receive a sentence already prepared for thought. It receives a sequence of tokens passing through computations that behave differently with respect to memory, transfer, and parallelism.

During prefill, the incoming tokens are processed to build attention representations. Iterative generation, or decode, follows: one token is produced, then used in the context for the next. Depending on workload size and architecture, prefill tends toward parallelizable compute intensity. Decode may collide with memory bandwidth, movement of weights, and conversation state. A chip that excels at the first phase need not offer the best final cost in the second. Tools such as TensorRT-LLM therefore document architectures that can disaggregate the resources used for these phases when needed, rather than forcing both into a single operating rhythm [S07](#).

A conversation also has a working memory unlike long-term storage. The key-value cache (KV cache) retains useful representations so parts of the context do not have to be recomputed. Every long conversation requires additional space. If a large block is reserved and only a small part is used, expensive chip memory may fill with empty space that nobody can bill. Look at memory as a librarian looks at shelves and the problem becomes clear: it is not always a shortage of books. It may be how the drawers were reserved.

In 2023, researchers published vLLM and its PagedAttention algorithm, inspired by virtual-memory paging in operating systems. Instead of wasteful contiguous allocation, they divided the attention cache into blocks that could be managed and shared flexibly. In the paper's peer-reviewed experiments, inference throughput improved by roughly two to four times against specified baseline systems at comparable latency [S05](#). That does not mean every data center will receive the same improvement by installing a piece of software. It does establish that an **algorithmic decision** can redistribute the value of existing hardware.

Before that, the Orca system presented at OSDI 2022 showed that iteration-level scheduling can reduce the time requests spend waiting behind longer batches [S06](#). Continuous batching, prefix caching, and disaggregating the computational phases are not decorative additions. They change how much work becomes a service within an acceptable time. Quantization reduces numerical precision and its memory and computational footprint, but it requires quality tests because the user may pay for the savings through a less accurate answer. Speculative decoding proposes tokens with a faster function and has another model verify them, so its speedup is not constant across workloads.

The theoretical peak of floating-point operations per second (FLOPS) must therefore not be equated with the number of correct answers a customer buys in an hour. Between the two stand a queue, a model, a scheduler, energy, a response-time contract, and questions of varying length. **Efficiency is not a fixed property of a chip. It is the behavior of an entire system performing a particular job.**

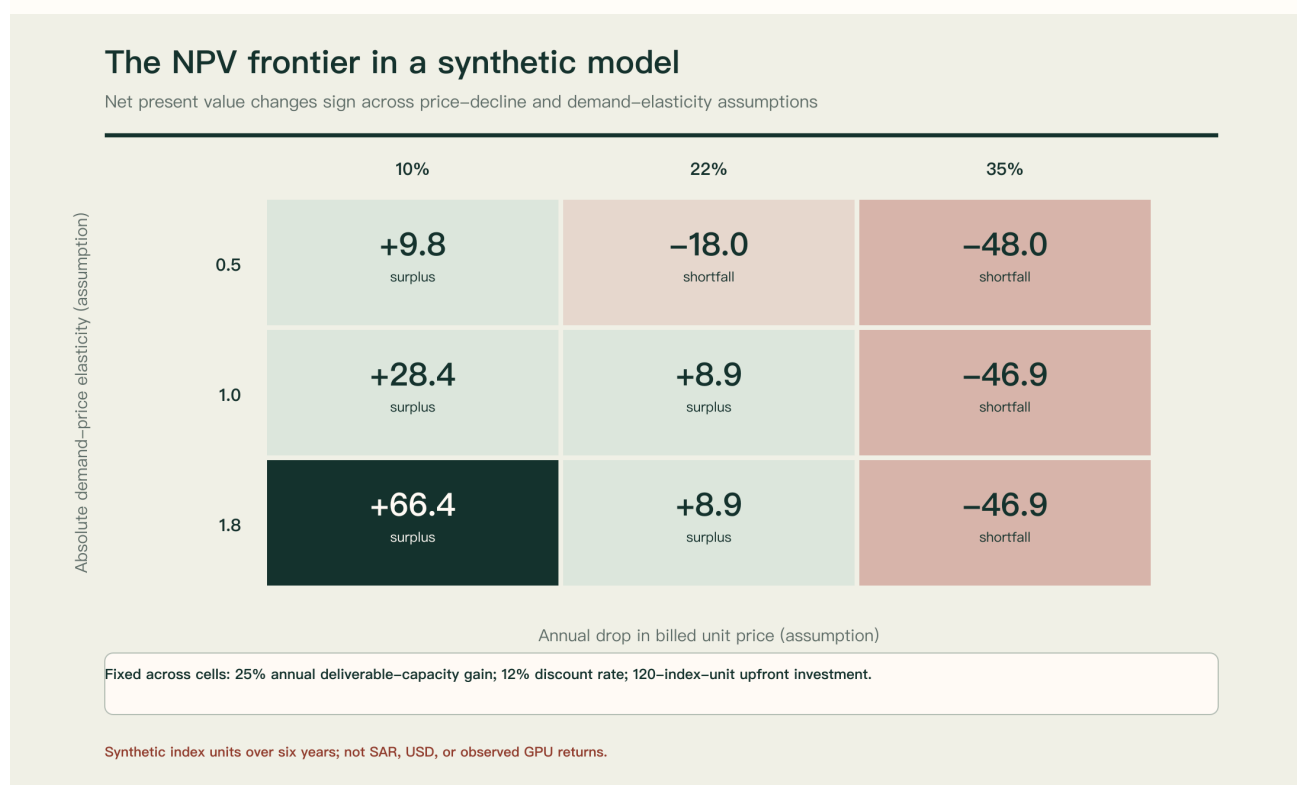
4. The paradox: software can rescue the investment and compress its price at the same time

Consider a hypothetical example, whose numbers should not be attributed to Saudi Arabia or any company. A server once completed fifty thousand work units a day. Better memory management raises its capacity to eighty thousand without replacing it. That is a welcome productivity gain. But if every competitor can deploy the same improvement, computation becomes less scarce and prices enter a new negotiation. The user benefits twice, first from speed and then from competition. The server owner is not guaranteed both benefits.

The outcome depends on what economists call price elasticity of demand: how much does the quantity demanded rise when price falls? In a simplified form, if demand is proportional to price raised to the negative power ε , revenue before capacity and cost constraints is proportional to price raised to $1-\varepsilon$. When ε exceeds one, total revenue may rise as price falls. When it is less than one, revenue may decline despite higher consumption. When capacity is full, mathematical enthusiasm about demand is irrelevant if its paid units cannot be delivered.

This is a precise version of the question William Stanley Jevons posed in *The Coal Question* in 1865: can efficiency increase total resource consumption rather than reduce it? He was not predicting units of artificial intelligence. More importantly, his historical observation does not license the slogan that every improvement must double demand. The **rebound effect** is a possibility determined by elasticity, available capacity, the price of alternative applications, and the value of the completed job. If history offers one rule here, it is this: do not infer a demand curve from an engineer's enthusiasm for a processor S08.

The market therefore deserves two opposing questions rather than one proclamation. Does software defer the need to buy the next generation, raising the return on existing equipment? Does that software also make the service abundant enough to reduce everybody's selling price? Both answers can be yes. The winner is determined by the cash margin left after the tokens are sold, not by the number emitted by the server.



Return map by elasticity and price decline, an illustrative simulation rather than company data

Figure 1. Net present value (NPV), measured in hypothetical units rather than dollars or riyals, for three degrees of price decline and three demand elasticities, assuming a 25% annual improvement in saleable capacity. Every cell is produced by running the accompanying model; none is a market forecast.

5. When utilization corrupts the financial picture

Investors sometimes look at GPU utilization as a hotel owner looks at occupied rooms. The analogy is useful until we remember that a hotel room does not turn from a suite into a chair in the corridor overnight because of a software update. Nor is all processor activity billable computation. It may be consumed by data preparation, a failed retry, waiting between network and memory, internal work the customer does not buy, or execution whose quality is too low to deliver.

A processor may appear 95% utilized yet earn less revenue than one at 65% because the price, the value of its requests, the amount of unbilled waste, or the service level differs. Conversely, utilization may fall after a successful software improvement because real capacity grew faster than demand. A lower percentage is not necessarily operational failure. This is a question about the denominator, not a moral tale about idleness and effort.

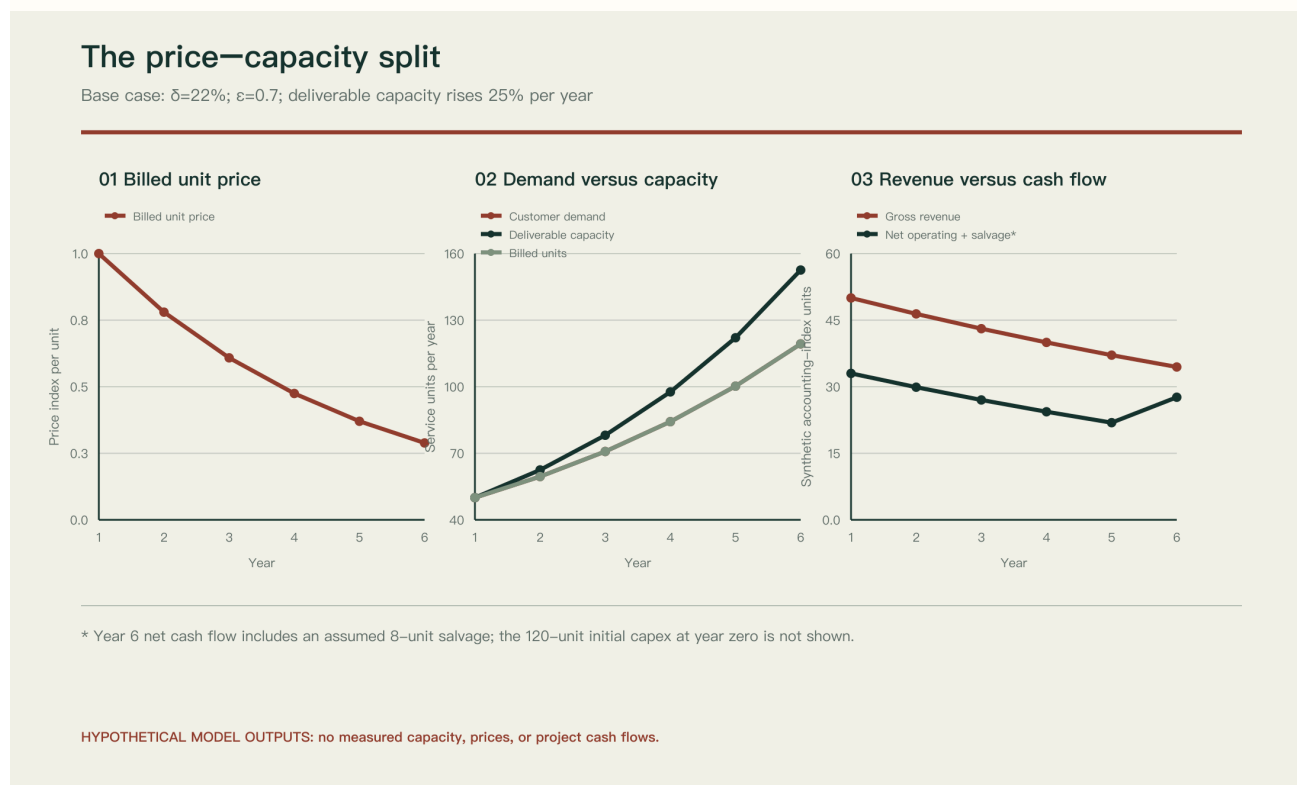
In the accompanying computational model, hypothetical capacity begins at fifty service units a year and unit price at one. We assume capital cost of 120 accounting units, a six-year window, a 12% annual discount rate, and disclosed fixed and variable operating costs. These values do not come from any company's accounts. Their value lies in making intuition falsifiable in a way rhetoric cannot.

When the assumed price falls 22% each year, effective capacity improves by 25%, and demand elasticity is **0.5**, NPV is **negative by about 18 units**. At an elasticity of **1.0**, it becomes **positive by about 8.9 units**. Raising elasticity to **1.8** does not increase the result under this pair of assumptions: it remains about **8.9**, because both demand paths strike the ceiling of deliverable capacity S12. A crowd of demand outside the gates does not pay for units the center cannot supply.

We ran a grid of 27 scenarios combining three rates of price decline, three elasticities, and three rates of capacity improvement. **Twelve produced positive NPV and fifteen negative NPV**. This is not a statistic about the probability of profit; every cell is a synthetic test case selected in advance. The stronger lesson is that both the appealing claim “cheap intelligence will consume the return on chips” and its opposite, “rising demand will save every investment,” fail as general laws. Each requires a specific measurable condition S12.

There is another trick more elegant than utilization. In the simulation's base case, the cash flows recover the total capital cost **in nominal terms in year five**, yet NPV remains **negative by about 5.6 units**. The arithmetic is not contradictory. When capital has a price, value returning after five years is not equal to value spent today. A project can celebrate “payback” while remaining less attractive than its alternative under a disclosed discount criterion. A picture of a long contract cannot answer the cost of waiting.

Our optimistic assumption must not be hidden: we allowed variable cost per unit to fall in step with capacity improvement. When a separate sensitivity test held that cost constant and left the other base-case assumptions unchanged, NPV fell from **negative 5.6** to **negative 31.2 units**. A polished slide about “higher productivity” cannot replace measurement of actual energy and operations per service. Completing more work faster does not always make each job cheaper S12.



Movement in price, volume, and revenue over six years, a hypothetical example

Figure 2. Three series from a single synthetic financial experiment. The physical movement of capacity diverges from the commercial movement of revenue; demand represents what the market wishes to buy, not what is actually delivered. The chart's limits and parameters are given in full in the research attachments.

6. Saudi Arabia: from a project of racks to an economy of real customers

On 31 August 2026, AMD, Cisco, and HUMAIN announced that computing infrastructure in Saudi Arabia built on AMD Instinct MI355X processors, Cisco Silicon One networking, and ROCm software **had entered service and was serving customers**. The same parties announced plans to begin deploying up to **250 megawatts** of capacity in 2027 and an ambition to expand to **one gigawatt** by 2030. These are three different states and must not be written as one: present operation according to the project owners' announcement, planned future expansion, and a distant goal dependent on execution, demand, and financing S09.

This news deepens the Saudi question. The presence of AMD processors in potentially paid service, rather than a discussion confined to buying from NVIDIA, opens a comparison of total cost of ownership (TCO), software compatibility, model portability cost, and actual application performance across hardware families. It does not prove that one platform is cheaper or more profitable than another, and it does not disclose customer contracts, utilization levels, or net margins. Those require data absent from the announcement.

A year earlier, HUMAIN announced the launch of an Arabic conversational application based on ALLaM 34B, with features including Arabic voice input, search, and the use of Arabic and English. The announcement does not say how many tokens were served or what revenue each session earned. It does move the image of the computing center from a model of electrical capacity to a possible use on a citizen's phone or inside an organization. That is the decisive economic transition. There is no market for computers merely because they think; there is a market because a person or an organization found work worth entrusting to them and paying for S10.

If Saudi Arabia wants to move from buying infrastructure to holding a competitive position in the inference industry, it need not choose between “the largest center” and “the cheapest token” as though either ended the inquiry. It must discover the kinds of high-value work it can deliver at predictable quality and a cost capable of declining over time. They might include pipelines developed for Arabic, specialized models with accuracy tests, highly private processing hosting, or model-development services with guaranteed response times. These are examples of economic categories within computing, not claims that they already exist or are all profitable.

7. One test can make five similar processors entirely different

Technical announcements have a charming way of compressing a large difference into one word: “performance.” Yet an inference engine serving five short questions concurrently is not the same engine when it carries a thousand long conversations sharing context. Time to first token (TTFT) differs from inter-token latency (ITL), and both differ from throughput. A device may win the token contest but answer late and lose an application that values immediacy. Another may rank lower in single-request speed but win under high concurrency.

On 16 September 2026, MLCommons released MLPerf Inference v6.1. It included new benchmarks closer to multistep tasks and results for new systems reviewed under the benchmark rules. This adds useful knowledge, but it does not grant any manufacturer a certificate of financial profitability. The benchmark measures operating performance under specified rules, not the price a computing operator realizes from customers, its financing, tax, or the cost of equipment sitting idle S11.

Imagine a real contest between two devices from the same generation, with no winner yet known:

- The first produces four thousand tokens per second under a benchmark workload, but its monthly operating cost per customer is high because it must reserve substantial capacity to guarantee time to first token.
- The second produces only three thousand five hundred in that workload, but shares contextual memory, runs heterogeneous batches efficiently, and benefits from users' repeated requests. It can therefore serve more customers within the time commitment for which they pay.

These are **invented illustrative numbers** showing how the comparison works, not measurements of two published devices. A winner cannot be declared until the intelligence model, input and output lengths, answer quality, actual request proportions, latency limits, electricity, cooling and operating costs, and billable share of work are held constant. Then “cost per accepted task” or “margin per SLA-compliant task” is more closely connected to economic life than a comparison of devices through isolated calculations.

The question now leaves the processor boxes and enters the investment director's office: which collection of workloads is suitable for this device at all? If customers require long sequential generation, memory and its bandwidth become constraints. If most usage searches shared corporate documents, context reuse may matter more. If customers need new models trained, the allocation of value and sensitivity to changing generations differ. It is not intelligent to put all these bills into a calculator labeled “hourly price” and then act surprised when the results diverge.

8. The trick that cannot repair a power contract with a sales contract

Three speeds in computing are difficult to place on a single dashboard: the speed of manufacturing a new generation, the speed of installing electrical capacity, and the speed of discovering demand willing to pay for the result. The first moves with product and research cycles. The second moves with construction, grids, and supply. The third moves through organizational experiments that may begin with enthusiasm and discover that inference is not needed all the time. This gap creates the value of a **real option**: the ability to delay a shipment, change the workload mix, or rent capacity in stages instead of turning every future ambition into an immediate bill.

Consider the contract whose image does not appear in data-center announcements. CoreWeave says many of its multiyear agreements are take-or-pay, and that the weighted average term of its committed contracts was about five years at the end of 2025. Under the

contracts it describes, the customer pays for committed capacity according to the terms even if it does not use all of it. This economic mechanism can shift some utilization risk away from the operator. It would therefore be unscientific to predict the effect of lower AI API prices on its revenue as though every service were sold immediately at an open spot price \$01.

A contract is nevertheless neither magical armor nor a game that erases time. Its effect depends on the company's ability to deliver, customer rights, pricing and renewal terms, the parties' creditworthiness, and debt payment dates. If a large customer loses its economic need for reserved capacity, it may continue paying until expiry and negotiate aggressively at renewal. If a computing center finds a profitable opportunity at a higher price, it may be unable to move capacity already promised elsewhere. The margin supplied by continuity can become an opportunity cost when conditions reverse. We do not have the details of Saudi contracts, so an American company's equation cannot automatically be projected onto them.

The secret to judging processor returns lies in the timing of surpluses and shortfalls, not merely the sum of cash flows. A loan coming due before demand matures can create cash pressure even if NPV becomes positive later. A long fixed-price contract may protect cash flows from a price decline while preventing full participation in an increase. A strong software improvement can raise unit margin and lower unit revenue at once. These exchanges should not be turned into a list of "governance risks." Each is a financial relationship that can be calculated: a repayment schedule, a demand curve, a consumption limit, a technology-generation date, and a window of electrical capacity.

9. A strategic idea: own the processor's economic learning curve, not the object

The greatest asset held by a computing operator in 2030 may be something the camera cannot see on a data-center tour: a precise history connecting **workload type** to what it actually consumed, what the customer paid for, and the quality the customer accepted. This is not a decorative log of call counts. It is knowledge learned month by month about where time was dissipated and where margin formed.

I call this research proposal the **Compute Commercial Life Map**. For each workload class, record five different clocks: the time required to prepare the model for the hardware; the wait before first output; the time to produce an acceptable output; the period for which the price remains competitively defensible; and the cash-payback period for that workload's share of the investment. These clocks are not equal, and none can be collapsed into a single national average. A chip's physical life may be six years, the life of its price advantage eight months, its useful life with a smaller model much longer, and the term of a contract occupying it unrelated to any of the three.

A machine vendor usually persuades you with the best benchmark. An experienced investment director wants an **Economic Advantage Decay Curve**: for how many months does the service remain more valuable than its alternative before it must lower its price or change the work it performs? That must be measured by equipment cohort, not the slogan attached to a processor generation. Three actions absent from the race's headline can then be compared: improving the software, repricing the service, or changing the customers suited to the capacity. Declining to buy an additional cohort may be the smartest computing decision of the year, even if it does not produce a glossy opening image.

There is an anti-pessimistic wager deserving respect. Cheaper inference may create applications that were previously impossible to sell, expanding the total market and strengthening profitability despite price declines. Complex new work may absorb every improvement instead of setting off a price war. Saudi Arabia may also benefit from energy availability, investment capacity, regional markets, and Arabic applications. These are conditions and test paths, not conclusions derivable from the announced size of a project. Until paid demand, unit margin, and executed contracts can be observed, both stories remain possible.

10. Five things worth reading before the next announcement

A researcher or investment official with authorized, aggregated operating and commercial data can begin tomorrow without waiting for a brilliant indicator that does not exist:

What we measure	The question it reveals	The error in relying on the easiest number
Revenue per Accepted Task	Does the customer pay more than the full cost of completing a successful task?	Relying on the price of a million tokens regardless of the task.
SLA-Constrained Throughput	How many requests can be completed at an agreed quality and latency?	Using the processor's laboratory maximum.
Contracted Revenue Share	What portion of cash flow is independent of spot price, and under what terms?	Equating a contract announcement with cash collection or guaranteed margin.
Price-Adjusted Payback	Does the project recover its capital while its commercial advantage survives?	Relying on accounting depreciation life.
Workload Margin Distribution	Does the average conceal workloads losing money and others subsidizing them?	Displaying an attractive aggregate utilization rate that obscures differences among applications.

Suppose an announced test takes place in the Saudi market. Two providers run comparable anonymized workloads and compete to deliver an Arabic service of known quality within the same time limit. Their full cost per accepted task is measured, then the experiment is repeated as demand volume, the distribution of conversation lengths, and the rate of context reuse change. Add an economic test: what does a 20% reduction in service price do to customers acquired, requests completed, and profit? There is no need to reveal details of critical facilities or customer secrets. Aggregated comparisons and measurement definitions are sufficient. The outcome is not guaranteed in advance. The newest device, the better software, the better-suited work, or the balanced contract may prevail.

This is the hypothesis I propose in this article: **“National advantage in AI computing arises not from owning the largest abstract capacity, but from the speed at which technical improvements are converted into useful paid work before their prices erode.”** It is a testable sentence. If, after controlling for workloads, costs, and contract terms, no relationship appears, it should be rejected or narrowed. I do not want to build a market from an elegant metaphor.

11. What will 2030 tell us, if we ask the right question today?

In one scenario, demand for complex inference expands faster than its price falls. The launch of a new hardware generation becomes an occasion to open applications that could not work on its predecessor, and Saudi capacity benefits from a larger regional market. In a second, software and new computers create enormous abundance, but most value moves to application developers and users; computing-center operators enter intense competition over price and cost. In a third, both sectors coexist: exceptional capacity for a limited class of expensive work and general-purpose computing that has become a low-margin commodity. These are **conditional paths with no probability assigned**, not three calculated predictions.

What should be observed between 2026 and 2030? Changes in realized prices, not published prices alone; the actual share of committed contracts; the composition of input, output, and cache; service performance under standardized time limits; differences in task cost across hardware generations on identical workloads; and the share of revenue from customers outside the domestic market. If token volume rises faster than its revenue, do not announce a disaster. Examine margin, then asset payback, then customer-acquisition cost. If revenue rises, do not announce victory. Examine how much of it capital, cooling, and networking consumed.

The old factory manager listened to the rhythm of the machines and knew that silence meant potential loss. The manager of a modern intelligence factory needs a different kind of courage: to watch the processor produce more, then ask, with great courtesy and greater severity, who

paid for that output, what work remained unbilled, and which customer could have used a smaller device for the same computation.

The modern machine may win every speed benchmark and lose the next pricing meeting. An older generation may lose its performance ranking and remain profitable because it performs work whose customers it knows well. There is no contradiction between the two. The contradiction is ours when we confuse scientific progress with progress on an invoice.

The question that should remain with the reader is not **how many processors will Saudi Arabia have?** It is **what will Saudi Arabia become better than others at computing, and who will need that computation enough to pay for it even after the next processor arrives?**

Sources and works cited

1. [CoreWeave, 2025 Form 10-K, SEC, useful life and contractual information, published 2026.](#)
2. [CoreWeave, Form 10-Q, period ending 30 June 2026, SEC, Note 2 and RPO.](#)
3. [NVIDIA, Rubin platform announcement, 5 January 2026.](#)
4. [OpenAI, Introducing GPT-6 Sol and Luna, September 2026, pricing and caching sections.](#)
5. [Kwon et al., Efficient Memory Management for Large Language Model Serving with PagedAttention, SOSP 2023, DOI 10.1145/3600006.3613165.](#)
6. [Yu et al., Orca: A Distributed Serving System for Transformer-Based Generative Models, OSDI 2022.](#)
7. [NVIDIA, TensorRT-LLM, Disaggregated Serving technical documentation.](#)
8. [William Stanley Jevons, The Coal Question, 1865, digitized historical edition.](#)
9. [AMD, Cisco and HUMAIN, joint production and expansion announcement, 31 August 2026.](#)
10. [Saudi Press Agency, HUMAIN Chat powered by ALLaM 34B, 25 August 2025.](#)
11. [MLCommons, MLPerf Inference v6.1 Results, 16 September 2026.](#)
12. [Reproducible original **synthetic** model, assumptions, and 27 scenario outputs; see the results table and assumptions.](#)

Editorial note: published financial figures are facts attributed to their sources; simulation results are not data about Saudi companies or services. The institutions named did not participate in preparing or reviewing this work.