

The token got cheaper. The task did not.

The usage-weighted price of a million LLM tokens has fallen by roughly half since the end of May 2026. In the same window, the sticker price of frontier reasoning went up, and Nvidia booked its largest data-centre quarter ever. These facts are not in tension: the market has split into a deflating commodity tier and an inflating premium tier, and buyers are trading up faster than prices are falling. The number that matters for the economy is not dollars per token — it is dollars per finished task, and that number is still rising.

AT

Alicante Tech Vibes

Research Labs

AUTHOR

~8 MIN READ · ALL FIGURES SOURCED & DATED

-52% USAGE-WEIGHTED TOKEN PRICE, 31 MAY → 3 SEP 2026	\$0.97 BLENDED COST PER MILLION TOKENS, 3 SEP 2026	+72% FRONTIER-TIER LIST PRICE INDEX, YEAR ON YEAR
4× GPT-5.6 SOL INPUT PRICE VS GPT-5 — FIRST FLAGSHIP RISE SINCE 2023	\$89.0B NVIDIA DATA-CENTRE REVENUE, Q2 FY27, UP 117% YOY	93% ENTERPRISES THAT OVERRAN THEIR AI BUDGET, JUL 2026 SURVEY

I. A price collapse nobody's invoice noticed

Silicon Data's LLM Token Expenditure Index — a daily, usage-weighted blend of what the market actually pays per million tokens across more than 200 open and

120 proprietary models, published on Bloomberg under the ticker SDLLMTK — printed \$0.97 on 3 September 2026. On 31 May it was \$2.04. In late July it was \$1.45, and by 6–8 August it had reached \$1.16–\$1.18, a level Jefferies flagged as the year's low in a note reported on 20 August. Taking the index at both ends gives a fall of about 52% in fourteen weeks (our arithmetic, from the two published values).

That is a genuinely large move in a genuinely short window, and it is the kind of number that gets read as a verdict on the whole AI trade. It should not be. **The index measures the average price of a token bought, not the price of any particular token, and most of its decline is composition, not repricing.**

FIG. 1: SDLLMTK USAGE-WEIGHTED PRICE, US\$ PER MILLION TOKENS



BARS SCALED TO THE 31 MAY MAXIMUM. MAY-AUGUST VALUES: JEFFERIES VIA SILICON DATA, REPORTED BY SCMP, 20 AUG 2026. SEPTEMBER VALUE: SILICON DATA, 3 SEP 2026.

II. Two indices, opposite directions

Run the same question through a different measurement and the answer inverts. BenchLM's Token Price Index, which tracks the median list price of models at each capability tier rather than weighting by usage, had the frontier tier at 16 in September 2026 against a March 2023 base of 100 — but **up 72% year on year**, with a median blended cost of \$6.00 per million tokens across 21 frontier models. The budget tier was up 36.8% on the year; only the mid tier fell, by 7.7%.

Both readings are correct. Frontier intelligence is 84% cheaper than it was in early 2023 and meaningfully more expensive than it was twelve months ago. The clearest single illustration came at the end of July: aggregators tracking OpenAI's published rates record GPT-5.6 Luna cut by 80% and Terra by 20% on 30 July, while the flagship Sol launched a day earlier at \$5 input and \$30 output per million — roughly four times GPT-5's \$1.25 and \$10, and the first increase in that product line's flagship price since March 2023. One vendor, one week, deflation and inflation in the same price list.

III. Who actually paid less

The honest technical read is that the deflation is real but narrowly distributed. [YipitData's June 2026 analysis](#), built on roughly 150 trillion tokens a month — near two quadrillion annualised — found effective pricing across API services down only 6% year-to-date in 2026, against a 39% fall in the second half of 2025. Split by customer, the divergence is stark: the developer market visible through OpenRouter fell about 15% through March, while enterprise consumption via cloud and Bedrock showed no meaningful change in effective price at all. Developers routed down-tier toward Kimi K2.5 and DeepSeek V4 Flash; enterprises routed *up*, toward the expensive models, despite cheaper substitutes being available.

Two mechanical caveats compound this. Prompt caching now cuts input costs by roughly 90% without any change in list price, so measured "prices" move for reasons unrelated to vendor pricing decisions. And index composition shifts as models enter and exit on sustained relevance — meaning a large SDLLMTK move can record the market's substitution behaviour rather than any seller's decision. Silicon Data itself has been explicit that the series is an expenditure-weighted average, not a price list.

Chief Product Officers should not confuse the deflation of commodity tokens with the democratization of frontier reasoning.

WILL SOMMER, GARTNER · CITED 20 JUN 2026. GARTNER ALSO FINDS AGENTIC WORKFLOWS CONSUME 5–30× MORE TOKENS PER TASK THAN CHATBOT QUERIES.

IV. Where the savings went: token intensity ate them

The demand side is what converts a price collapse into a bigger bill. [Gartner puts agentic workflows at five to thirty times the tokens per task](#) of a simple chatbot query; McKinsey's 2026 work goes further, putting multi-turn agentic workloads at up to a thousand times standard chat, with the final visible output accounting for only 5–15% of tokens consumed. EY's estimate of a single customer-service interaction moves from about \$0.04 in 2023 to about \$1.20 in

2026 — a thirty-fold rise in cost per completed task while the cost per token fell by orders of magnitude.

The budget consequences are already recorded. [McKinsey's July 2026 Enterprise AI FinOps survey](#) found [93% of respondents had exceeded their AI budgets](#), with a fifth limiting AI use on operating cost grounds. Reporting from late July put inference at roughly [85% of enterprise AI budgets](#), and cited Uber exhausting an annual AI budget by April after deploying a coding assistant to 5,000 engineers the previous December. This is Jevons in a quarterly reporting cycle: cheaper units, far more units, larger total.

FIG. 2: ESTIMATED COST OF ONE CUSTOMER-SERVICE INTERACTION, US\$

2023		\$0.04
2026		\$1.20

BARS SCALED TO THE 2026 VALUE. EY ESTIMATE, CITED 20 JUN 2026. PER-TOKEN PRICES FELL OVER THE SAME PERIOD; PER-TASK COST ROSE ABOUT THIRTY-FOLD.

V. The premium tier is the moat, and the open models know it

Academic work published in March 2026 anticipated exactly this split. Mingdeng Du of Wuhan University finds [economy-tier token prices halving every 1.10 years](#) and [mid-tier every 1.55 years](#), while the flagship tier resists [exponential decay entirely](#), carrying an average reasoning premium of 31.5×. The paper attributes essentially all cost reduction to software and architecture rather than GPU hardware, and records concentration falling sharply — a Herfindahl-Hirschman Index down 54.2%, from 4,558 to 2,086, crossing out of "highly concentrated" territory. Its own stated limits are real: 62 milestone observations and proxy-based market shares.

The open-weight challenge is closing the commodity gap and not yet the frontier one. [DeepSeek V4 left preview on 13 August 2026](#) at \$0.66/\$1.98 per million off-peak for the Pro tier, scoring 80.6% on SWE-bench Verified — competitive on the older benchmark, but roughly 55% on the harder SWE-bench Pro against 64.3% for Claude Opus 4.7, with a NIST CAISI evaluation placing it about eight months behind leading US models. That eight-month gap is, for now, what the reasoning premium is priced on.

VI. The supply side is not behaving like a deflating industry

If falling token prices signalled falling demand, the hardware would show it. It does not. [Nvidia's Q2 FY2027 results, released 26 August 2026](#), put revenue at \$96.2 billion, up 106% year on year, with data centre at \$89.0 billion, up 117%, gross margin at 75.0%, and third-quarter guidance of \$108.0 billion. Jensen Huang's framing was that compute is now revenue.

The bear case is that this is precisely the setup for a reckoning. Allianz Research has put the gap between AI capex growth and sales growth at 46%, against the 32% divergence that preceded the 2001 telecom bust, and macro strategist Andreas Steno Larsen argued in early July that sustained token-price weakness would end the memory, hardware and data-centre trades for this cycle. [Bloomberg's own coverage of that 20% drawdown was careful to note the ambiguity](#): a falling expenditure index is equally consistent with customers migrating to cheaper models as with demand weakening, and one index move is not a rerating. [Citadel Securities reached a similar conclusion on 10 June 2026](#), forecasting uneven diffusion: frontier capability concentrating among well-capitalised firms with hard problems, everyone else settling into the cheap tier.

That distributional point is the one with macro consequences. PwC's 2026 work found the top 20% of companies capturing 74% of AI's economic value, only 12% of CEOs reporting both cost and revenue benefits, and 56% reporting no significant financial benefit yet. Cheap tokens have not, so far, broadened the winners' circle.

VII. How it plays out: three scenarios

Horizons are two to five years. Probabilities are our analytical judgment, not measurements, and are stated to make the reasoning falsifiable rather than to convey precision.

1 · Baseline: the split holds ~55%

2026–2029 · STEADY

Commodity tokens keep deflating while the frontier holds its premium, and enterprise spend keeps climbing on volume. Watch for SDLLMTK continuing to fall while BenchLM's frontier tier stays flat or rises; for effective enterprise pricing in usage panels staying within single-digit annual declines; and for Nvidia landing near its \$108 billion guide without needing a demand narrative rescue. In this world the winners are inference-efficiency vendors, routing layers, and whoever owns the reasoning tier.

2 · Acceleration: the premium breaks ~25%

2027–2029 · FAST

An open-weight model closes the hard-benchmark gap and a frontier lab is forced to cut flagship prices, at which point deflation finally reaches enterprise invoices and task-level automation reprices across mid-market services. Leading indicators: an open model within a few points of the closed leaders on SWE-bench Pro rather than SWE-bench Verified; NIST CAISI's capability lag narrowing from eight months toward three; and the first list-price cut on a flagship reasoning tier since GPT-5.

3 · Stall: the index is read as demand ~20%

2026–2028 · ABRUPT

Markets treat falling token prices as evidence of weak demand, the capex-to-sales gap closes the hard way, and financing for the buildout tightens before the revenue arrives. Indicators: a hyperscaler cutting 2027 capex guidance; Nvidia guiding below consensus for the first time in the cycle; and — the decisive one — token *volume* growth decelerating at the same time as price, which would separate substitution from genuine demand weakness.

WHAT WOULD FALSIFY THIS THESIS

The claim is that unit deflation is being absorbed by token intensity and trade-up, so cost per completed task is not falling. It fails if effective enterprise pricing turns sharply negative — say a 25%+ decline over two consecutive quarters in a usage-weighted panel like YipitData's — while agentic token consumption per task plateaus rather than grows. It also fails if the frontier-tier inflation recorded by BenchLM proves to be a composition artefact of models entering and exiting the tier rather than real repricing; the GPT-5.6 Sol increase is one clean data point, not a trend. The strongest counter-case on the record is Steno Larsen's, reported by Bloomberg on 3 July 2026: that sustained token-price weakness is a demand signal, not a mix signal, and would end the hardware trade for this cycle. Current evidence does not settle it, because volume and price are falling and rising in ways the published indices cannot yet cleanly separate.

WHAT TO WATCH, IN ORDER

1. **SDLLMTK against token volume.** Daily on Bloomberg. Price falling with volume rising is substitution; both falling is demand weakness. Checkable now, decisive within two quarters.
2. **Nvidia Q3 FY2027, due late November 2026.** The \$108bn guide is the cycle's cleanest test of whether cheap tokens are shrinking or expanding the compute bill.
3. **The next flagship price move.** Any cut to a top reasoning tier would mark the premium breaking; another increase confirms the split. Watch OpenAI, Anthropic and Google list pages.
4. **Open-weight scores on hard benchmarks.** SWE-bench Pro and the next NIST CAISI evaluation, not SWE-bench Verified — the eight-month lag is what the premium is priced on.
5. **Hyperscaler 2027 capex guidance, Q4 earnings season, Jan–Feb 2027.** The first cut would move the Stall scenario from tail to base case.

SOURCES

1. Silicon Data, “LLM Token Expenditure Index (SDLLMTK),” value as of 3 Sep 2026 · silicondata.com

2. South China Morning Post, "Enterprise AI costs hit 2026 low driven by price wars, Chinese open-source models," 20 Aug 2026, citing Jefferies and Silicon Data · [scmp.com](https://www.scmp.com)
3. BenchLM, "Token Price Index," September 2026 reading · benchlm.ai
4. BenchLM, "OpenAI API Pricing," data as of 3 Sep 2026 · benchlm.ai
5. ValueAdd VC, "OpenAI API pricing 2026," citing OpenAI published rates as of 27 Aug 2026 · valueaddvc.com
6. YipitData, "What Can Nearly 2 Quadrillion Annualized Tokens Tell Us About LLM Pricing Trends?," June 2026, data through May 2026 · yipitdata.com
7. NeuralWired, "Gartner: AI Inference Cost Won't Drop Your Bill in 2026," 20 Jun 2026, citing Gartner and EY · neuralwired.com
8. Business Standard, "Tokens to usage: Why cheaper AI still leaves enterprises with bigger bills," 11 Aug 2026, citing McKinsey, PwC and the Stanford AI Index · business-standard.com
9. TechNewsWorld, "Falling Token Prices Fail To Slow Enterprise AI Spending," 28 Jul 2026 · technewsworld.com
10. Mingdeng Du (Wuhan University), "Tiered Super-Moore's Law: Price Evolution, Production Frontiers, and Market Competition in LLM Inference Services," arXiv, March 2026 · arxiv.org
11. The AI Rankings, "DeepSeek V4: Benchmarks, Pricing & Specs," GA 13 Aug 2026, citing NIST CAISI · theairankings.com
12. NVIDIA, "NVIDIA Announces Financial Results for Second Quarter Fiscal 2027," 26 Aug 2026 · nvidianews.nvidia.com
13. AI Weekly, summarising Bloomberg, "The AI Trade Is Losing One of Its Key Signals," 3 Jul 2026, citing Andreas Steno Larsen and Allianz Research · aiweekly.co
14. Citadel Securities, Frank Flight, "Tokenomics," 10 Jun 2026 · citadelsecurities.com

METHOD NOTE: FIGURES ARE ATTRIBUTED AND DATED INLINE; WHERE INDEPENDENT SOURCES DISAGREE, THE DISAGREEMENT IS REPORTED RATHER THAN AVERAGED. THE -52% HEADLINE IS COMPUTED FROM TWO PUBLISHED INDEX VALUES, NOT A PUBLISHED FIGURE. SCENARIO PROBABILITIES ARE ANALYTICAL JUDGMENT.

Built by the **Alicante Tech Vibes** community, a tope.

2026-09-05