The price of intelligence fell 43% in ten weeks. Almost none of it was a price cut.

Between 31 May and 8 August 2026, the blended price the market actually pays for a million LLM tokens fell from $2.04 to $1.16. Macro desks are reading that line as a demand signal — the moment AI spending finally rolled over. It is mostly something else: buyers substituting into open-weight models, most of them Chinese, faster than any vendor cut list prices. The distinction decides whether $725–785bn of 2026 hyperscaler capex is early or wasted.

AUTHOR

AT

Alicante Tech Vibes *Research Labs*

~8 MIN READ · ALL FIGURES SOURCED & DATED

–43% BLENDED TOKEN PRICE, 31 MAY → 8 AUG 2026	**$1.16** INDEX LOW PER 1M TOKENS, 6–8 AUG 2026	**72.4%** OPEN-WEIGHT SHARE OF TOP-10 ROUTED TOKENS, LATE JUL 2026
$725–785bn 2026 HYPERSCALER CAPEX, +77% YEAR ON YEAR	**5–30×** TOKENS PER AGENTIC TASK VS. ONE CHAT TURN (GARTNER)	**$0.14** DEEPSEEK V4-FLASH-0731 INPUT, MIT LICENCE, 31 JUL 2026

I. Ten weeks, forty-three percent

On 10 August, Jefferies circulated a note flagging that enterprise inference costs had hit their 2026 low, citing the Silicon Data LLM Token Expenditure Index — a daily, expenditure-weighted measure of what the market pays per million tokens across frontier APIs, open-weight platforms, brokered instances and self-hosted deployments. The series read $2.04 on 31 May, $1.45 in late July, and $1.16–$1.18 between 6 and 8 August. That is a 43% decline in ten weeks in the realised price of the core input to every AI product on the market.

Two dated events sit inside the steepest part of the curve. On **30 July**, OpenAI cut GPT-5.6 Luna from $1/$6 to $0.20/$1.20 per million input/output tokens — roughly 80% off — and Terra by 20% to $2/$12, leaving the flagship Sol tier untouched at $5/$30. The company attributed the cut to rewritten production GPU kernels and a redesigned speculative-decoding draft model, framing it as an efficiency dividend rather than a loss leader. On **31 July**, DeepSeek shipped V4-Flash-0731, a 284B-parameter mixture-of-experts model with 13B active parameters under an MIT licence, and held its price at $0.14/$0.28 while improving agentic benchmark scores.

FIG. 1: BLENDED COST OF ONE MILLION LLM TOKENS, EXPENDITURE-WEIGHTED

BARS SCALED TO THE 31 MAY READING (=100%). SILICON DATA LLM TOKEN EXPENDITURE INDEX, VIA JEFFERIES RESEARCH REPORTED 10 AUGUST 2026.

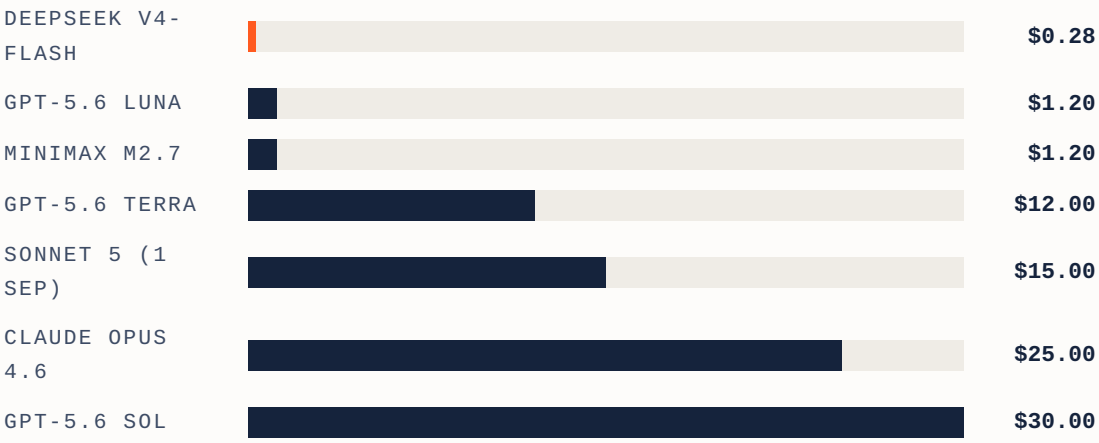
II. The buyers moved before the sellers did

An expenditure-weighted index does not measure list prices. It measures what buyers chose. And by mid-2026 buyers had already moved: the State of Open Source AI report (v1.0.1, July 2026) finds the seven highest-volume models on OpenRouter all ship open weights, with open models taking **72.4% of top-10 routed token volume by late July** — up from roughly a third in late 2025. Chinese open-weight providers served about 46% of routed tokens against 35% from US providers.

The composition matters more than the headline. A separate Q2 2026 breakdown puts Xiaomi at 21.1% of weekly OpenRouter tokens, Alibaba's Qwen at 13.9%, and MiniMax, Zhipu, DeepSeek and StepFun together at 24.6% — a

Chinese share that stood below 2% a year earlier. On coding volume the capability gap has effectively closed; on reasoning benchmarks closed models still lead by three to eight points, and on multimodal and computer-use tasks the lead is wider. Buyers are not abandoning frontier models. They are routing the cheap 90% of their traffic away from them, which drags a spend-weighted average down without anyone cutting a price.

FIG. 2: COST OF ONE MILLION OUTPUT TOKENS, SELECTED MODELS, AUGUST 2026



BARS SCALED TO $30 (=100%); THE CHEAPEST BARS ARE SLIVERS BY DESIGN — THE SPREAD IS THE FINDING. SOURCES: ORCA ROUTER (30 JUL 2026), DIGITAL APPLIED PRICING TRACKER (AUG 2026), HUGGING FACE MODEL CARD (31 JUL 2026).

III. What an expenditure-weighted index cannot tell you

The honest read is that this index is being asked to carry more than it can. CIO reported on 3 July that the index had fallen 20% from its May peak and that the weighting methodology makes the cause genuinely unidentifiable: enterprise price pressure, a shift to less token-hungry models, and outright demand weakness all produce the same line. Silicon Data's own reading, reported on 8 June, was that the softness reflected slowing migration toward premium closed models rather than a reversal — while macro strategist Andreas Steno Larsen warned in the same piece that sustained weakness would end the memory, hardware and data-centre trades for this cycle. Both cannot be right, and the index alone cannot adjudicate between them, because Silicon Data publishes no matching volume series.

There is a harder caveat: **not every price is falling.** Anthropic's Claude Sonnet 5 is running a promotional $2/$10 through 31 August that reverts to $3/$15 on 1 September — a 50% increase. DeepSeek has announced a peak-hours

surcharge doubling all billing items during two daily Beijing windows, with no effective date set, and has signalled a further increase it calls significant. Part of the August low is therefore a promotional artefact with a known expiry.

The price of a token fell 43% in ten weeks. The price of a finished task did not.

GARTNER PUTS AGENTIC WORKLOADS AT 5–30× THE TOKENS OF A CHAT TURN; STANFORD WORK CITED ALONGSIDE IT FINDS CODING AGENTS REACHING FAR HIGHER MULTIPLES, WITH 30× VARIANCE BETWEEN IDENTICAL RUNS OF THE SAME TASK. SPHERON, 2026.

IV. The capex math now runs on elasticity

Microsoft, Google, Amazon and Meta have guided to a combined $725–785bn of 2026 capital expenditure, roughly 77% above 2025 and an estimated 93% of operating cash flow, against 33% in 2023. That analysis puts the capex-to-revenue divergence near 46%, already wider than the 32% peak of the 2001 telecom overbuild. An earlier February tally at $660–690bn flagged the same structural risk: infrastructure built today may take 18–36 months to earn proportionally, and Amazon's stock fell 8–10% on its announcement.

That investment is underwritten by revenue billed per token. If realised price per token falls 43% while token volume rises faster, the buildout is early rather than wrong. The volume evidence is real: OpenRouter alone reached 25 trillion tokens a week, about 100 trillion a month, a fivefold rise in six months as of 26 May. Jefferies made the Jevons argument explicitly on 26 June, seeing zero sign of capex slowing and rotating into memory — SK Hynix, Kioxia, Samsung — while trimming internet names. *Judgment, not fact:* elasticity above one is the load-bearing assumption of the entire 2026 capex cycle, and it is currently supported by platform-level volume data rather than audited industry figures.

V. Cheap tokens, expensive bills, invisible margins

For buyers, the 43% decline has not shown up as savings. The arithmetic is that cost equals price per token multiplied by tokens per task, and the second term has grown faster: one analysis puts the fall in token prices since 2023 at roughly a thousandfold against a consumption rise several orders larger, driven by

orchestration overhead, context accumulation, retry loops and tool descriptions. Worked examples show a fraud-check agent at 13,500 tokens against 800 for a chat turn, and one fintech tripling its bill while growing users tenfold — cheaper per unit, more expensive in total.

The supply side has its own blind spot. Fireworks, Baseten and Together AI raised roughly $3.8bn in four weeks selling managed inference on customer-tuned open weights, yet as Forbes noted on 18 July, none disclose gross margins — the one number that would reveal whether falling token prices are an efficiency dividend or a subsidy. Long GPU commitments against short customer contracts leave little room to hold a price through a supply shock. Morgan Stanley's 3 August framing is the useful one: enterprise AI tasks average around $55 of revenue against $2–5 of direct cost, 63% of surveyed enterprises already run open models, and moving a workload from closed to open cuts price roughly 70%. At those spreads, substitution is close to free money for the buyer and a direct revenue transfer away from the frontier labs.

VI. How it plays out: three scenarios

Horizons are two to five years; probabilities are analytical judgment, not model output, and reflect the balance of evidence as of 12 August 2026.

1 · Baseline — mixed architecture holds ~55% 2026–2028 · STEADY

The index stabilises between $1.00 and $1.40 as the September promotional reversions offset continued substitution; open weights take the high-frequency tier, frontier models keep reasoning and multimodal work, and capex guidance holds. Watch for the index recovering less than 15% after 1 September, hyperscaler AI revenue growth staying above 60% year on year in Q3 reporting, and no downward revision to 2027 capex.

2 · Acceleration — elasticity wins ~25% 2026–2029 · FAST

Sub-$0.50 blended pricing makes always-on agent workloads viable and volume growth outruns the price decline decisively, validating the Jefferies memory trade. Watch for OpenRouter weekly volume clearing 50 trillion tokens, DRAM and HBM contract prices rising into Q4, and at least one hyperscaler raising 2027 capex guidance at October earnings.

3 · Stall — it was demand after all ~20%

2026–2027 · ABRUPT

Pilots fail to convert, token volume flattens, and the index keeps falling on weakness rather than substitution — the Steno Larsen case, which would break the memory, hardware and data-centre trades together. Watch for the index below $0.90 alongside flat routed volume, cloud AI revenue growth decelerating below 40%, and any hyperscaler trimming capex guidance.

WHAT WOULD FALSIFY THIS THESIS

The argument here is that the 43% decline is substitution, not contraction. It fails if Silicon Data — or any independent source — publishes a routed-volume series showing flat or falling token consumption across the same 31 May to 8 August window; in that case the index is a demand signal and the bearish reading is correct. It also fails, more mundanely, if the index rebounds above $1.80 by mid-October once Anthropic's promotion expires on 1 September and DeepSeek's surcharge activates, which would mean the August low was a promotional artefact rather than a structural repricing. The strongest counter-case on the record is CIO's 3 July caution that the index's weighting makes any causal claim, including this one, unidentifiable from the published series alone. No volume series is currently public, so this thesis is not yet testable against its own key datum.

WHAT TO WATCH, IN ORDER

1. **1 September 2026** — Claude Sonnet 5 reverts from $2/$10 to $3/$15. The size of the index response over the following fortnight separates promotional noise from structural decline.
2. **Late August 2026** — Nvidia's next quarterly report and its memory and data-centre commentary, the cleanest read on whether cheap tokens are pulling hardware demand up or down.
3. **Undated, announced** — activation of DeepSeek's 2× peak-hours surcharge, which would be the first structural price rise from a leading open-weight provider and a signal that inference capacity, not model quality, has become the binding constraint.
4. **October 2026** — Q3 hyperscaler earnings: whether 2027 capex guidance rises, holds or is trimmed against the 46% capex-to-revenue divergence.
5. **Q4 2026** — OpenRouter weekly token volume against the 25 trillion May baseline, the closest public proxy for whether elasticity exceeds one.

SOURCES

1. Silicon Data, “LLM Token Expenditure Index (SDLLMTK)” · silicondata.com
2. South China Morning Post, “Enterprise AI costs hit 2026 low driven by price wars, Chinese open-source models,” 10 August 2026 · scmp.com
3. CIO, “AI token prices are cooling — but why?” 3 July 2026 · cio.com
4. Let’s Data Science, “Silicon Data Index Signals AI Spend Inflection,” 8 June 2026 · letsdatascience.com
5. Orca Router, “OpenAI Cuts GPT-5.6 API Prices,” 30 July 2026 · orcarouter.ai
6. Digital Applied, “AI API Pricing, August 2026: Cuts, Promos, and Traps,” August 2026 · digitalapplied.com
7. Hugging Face, “DeepSeek V4 Flash Is Now Official: What Changed in the 0731 Build,” 31 July 2026 · huggingface.co
8. XenoSpectrum, “DeepSeek V4 Flash 0731 Launches Public Beta with Price Held at $0.14,” August 2026 · xenospectrum.com
9. State of Open Source AI, v1.0.1, July 2026 · stateofopensource.ai
10. Digital Applied, “Open-Weight vs Closed-Source AI Models 2026: Gap Analysis,” Q2 2026 · digitalapplied.com
11. Forbes, Janakiram MSV, “Open Weight Models Are Turning Inference Into A Control Point,” 18 July 2026 · forbes.com

12. Tech Startups, “OpenRouter raises $113M as AI token usage surges to 100 trillion monthly,” 26 May 2026 · techstartups.com
13. NextWaves Insight, “What Q2 2026 Earnings Must Show on AI Capex ROI,” July 2026 · nextwavesinsight.com
14. Futurum Group, “AI Capex 2026: The $690B Infrastructure Sprint,” 12 February 2026 · futurumgroup.com
15. Let’s Data Science, “Jefferies Says Cheaper AI Models Boost Infrastructure Demand,” 26 June 2026 · letsdatascience.com
16. WEEX, “Will Open Source Models Kill Computing Demand? Morgan Stanley Explores Three Futures of AI,” 3 August 2026 · weex.com
17. Spheron, “Agentic AI Inference Cost: Why Agents Burn 5–30x Tokens,” 2026 · spheron.network
18. The Modern Data Company, “Why Cheaper AI Tokens Are Increasing Enterprise AI Costs,” 2026 · themoderndatacompany.com

METHOD NOTE: FIGURES ARE ATTRIBUTED AND DATED INLINE; WHERE INDEPENDENT SOURCES DISAGREE, THE DISAGREEMENT IS REPORTED RATHER THAN AVERAGED. SCENARIO PROBABILITIES ARE ANALYTICAL JUDGMENT. VALUATION AND RUN-RATE FIGURES CITED FROM PRIVATE COMPANIES ARE COMPANY-SUPPLIED AND UNAUDITED.

Built by the Alicante Tech Vibes community, a tope. 2026-08-12