

The frontier stopped getting cheaper

In the first two weeks of September 2026, OpenAI and Anthropic released new flagship models at exactly the same list price: ten dollars per million input tokens, fifty per million output. Below them, the commodity tier kept collapsing — Gemini 3.8 Flash at \$0.75, DeepSeek V4.1-Flash at \$0.15 off-peak. The deflation that enterprise AI budgets were built on is still real, but it has stopped happening at the tier where agents actually run. For anyone modelling AI cost over the next three years, the per-token price is now the wrong number to forecast.

AT

Alicante Tech Vibes

Research Labs

~9 MIN READ

·

ALL FIGURES SOURCED & DATED

\$10 / \$50 IDENTICAL LIST PRICE PER 1M TOKENS FOR GPT-6 ASTRA AND CLAUDE FABLE 5.1, SEPT 2026	83× ASTRA’S OUTPUT PRICE VS DEEPSEEK V4.1-FLASH OFF-PEAK (\$50 VS \$0.60)	0% MONTH-OVER-MONTH MOVE IN THE FRONTIER PRICE INDEX, 11 SEPT 2026
12× / 300× FRONTIER VS BUDGET-TIER PRICE DECLINE, MARCH 2023 TO 2026	5×+ FORECAST RISE IN INFERENCE COST PER AGENTIC WORKFLOW THROUGH 2028 (GARTNER)	272K INPUT TOKENS ABOVE WHICH ASTRA REPRICES THE WHOLE REQUEST AT 2× INPUT

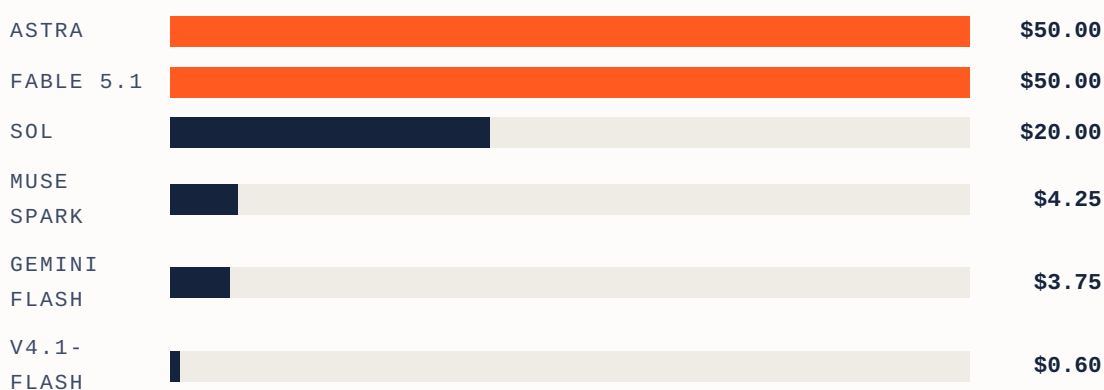
I. Two launches, one price

On 1 September 2026 Anthropic shipped Claude Fable 5.1 and Mythos 5.1 at \$10 per million input tokens and \$50 per million output. Two or three days later — sources differ on whether it was the 3rd or the 4th — OpenAI shipped GPT-6 Astra at \$10 and \$50. Not close. The same.

That is a new fact, and a narrow one, so it is worth stating precisely what it is and is not. It is not evidence of collusion; identical prices emerge routinely in markets with two credible sellers, similar cost structures and public price lists. It is evidence that **neither lab believes it can win frontier customers by undercutting the other**, and that both now expect to compete on capability and on the structure of the bill rather than on the headline rate. OpenAI, as one pricing review put it, stopped competing on price at the top.

The gap to everything below is the part that should reorganise a budget. Google's Gemini 3.8 Flash, out 2 September, lists at \$0.75 / \$3.75. Meta's Muse Spark 1.3 came in at \$1.25 / \$4.25. DeepSeek's V4.1-Flash, released 11 September with a 1M-token window and a 552-billion-parameter mixture-of-experts design, charges \$0.15 / \$0.60 off-peak. The market has not converged on a price. It has split into two markets.

FIG. 1: OUTPUT PRICE PER 1M TOKENS, STANDARD TIER, MODELS RELEASED 1-11 SEPTEMBER 2026



BARS SCALED TO MAX (\$50). SOL AND GEMINI FLASH FIGURES ARE PROMOTIONAL RATES WITH STATED EXPIRY DATES; V4.1-FLASH IS THE OFF-PEAK RATE. SOURCES: ANTHROPIC, LLM-STATS, CLOUDZERO, GCN, DATACONOMY, YAHOO FINANCE, SEPT 2026.

II. Competition moved into the fine print

Price competition did not stop. It moved to levers that are conditional, revocable and much harder to put in a spreadsheet.

Anthropic held Fable's base rate flat and cut cache reads from \$1.00 to \$0.25 — a 75% reduction that the company estimates makes typical workloads about 25% cheaper and heavily agentic ones up to roughly 45% cheaper. That is a real cut, and it only lands if your workload is cache-heavy. DeepSeek prices by clock: the \$0.15 / \$0.60 rate applies off-peak and doubles during peak windows. Google's \$0.75 / \$3.75 is an introductory rate that expires on 31 December 2026 and then doubles to \$1.50 / \$7.50. OpenAI's cheaper GPT-5.6 Sol tier is a promotional rate guaranteed only through 21 November 2026.

Astra adds surcharges in the other direction. Cross 272,000 input tokens and the entire request reprices at double the input rate and 1.5× output — a 280K-token request costs about \$7.10 where the same work trimmed to 272K costs roughly \$3.72, per CloudZero's worked example. Low-latency “fast” mode doubles the standard rate again. The headline is a 1.05-million-token context window; the affordable part of it ends at about a quarter of that.

III. The index is still falling, and that is the trap

The strongest case against this brief is that prices are plainly still dropping, and it deserves a fair hearing. BenchLM's frontier price index stood at 16 on 11 September 2026 against a March 2023 baseline of 100 — an 84% decline in three and a half years. Independent trackers put the fall further still when measured against capability rather than calendar.

But the aggregate hides a split that matters more than the average. TokenCost's price index, which tracks every major published price change since March 2023, separates the two: the cheapest model capable of a given task fell by roughly 200–300×, while the *best available* model fell only about 12×, from \$30 to \$2.50 per million. The widely quoted “300× cheaper” figure is a statement about the commodity tier, not the frontier. And momentum has now gone flat: BenchLM recorded a 0% month-over-month move in that same frontier index, cautioning that there is no guarantee of a straight-line decline.

A second caveat cuts against the thesis honestly. Astra is priced 2.5× Sol but appears to consume fewer tokens per agentic task, so part of the premium is recovered in efficiency. Per-token price was never the right unit; cost per resolved outcome is. That is exactly why the flat frontier list price is a warning rather than a curiosity — **the one number enterprises can forecast has**

stopped moving, and the numbers that replaced it are workload-dependent, promotional, or both.

“Each successive generation of AI capability will necessitate more, and often more expensive, tokens.”

WILL SOMMER, SR. DIRECTOR ANALYST, GARTNER · PRESS RELEASE, 17 AUGUST 2026, FORECASTING THAT INFERENCE COST PER AGENTIC WORKFLOW RISES MORE THAN FIVEFOLD THROUGH 2028.

IV. The bill lands on the workflow, not the token

Gartner’s August forecast is the clearest statement of the mechanism. Routing a task to an agentic reasoning model rather than a chatbot raises inference cost at least fivefold, and cost per agentic workflow is forecast to rise more than fivefold through 2028 even as token prices fall. Efficiency gains are being spent, not banked: cheaper tokens make more expensive models affordable, and more capable models are given longer-horizon work that consumes far more tokens.

Trace that through the obvious channels. On **productivity**, the arithmetic of automation no longer improves automatically with time — a task that is uneconomic at \$50 per million output tokens today is not reliably economic in eighteen months, so the pilot that failed its business case does not get rescued by waiting. On **market structure**, a stable frontier price and a collapsing commodity price push work down-tier: the rational architecture is a cheap model doing most of the volume, escalating to the frontier only where it demonstrably pays, which is precisely the “inference tiering” Gartner recommends. On **capital flows**, the frontier premium is what services the buildout — Futurum put 2026 hyperscaler capex at \$660–690 billion across five firms in its February analysis, nearly double 2025.

V. Why the floor is being defended now

The timing is not incidental. Anthropic is reported to have selected Nasdaq for a potential October listing at a valuation approaching \$2 trillion, alongside a \$50 billion US datacentre programme announced with Fluidstack in November 2025. SoftBank closed an upsized \$11.9 billion loan on 14 September to fund its

OpenAI position. A company weeks from a roadshow does not cut gross margin on its flagship product, and a lender syndicate does not underwrite a price war.

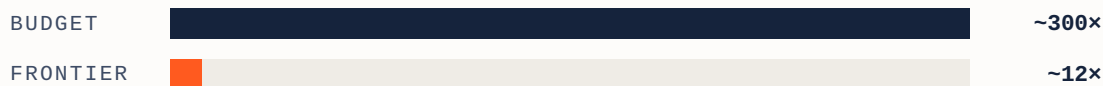
This is judgment rather than fact, and should be read as such: the identical \$10/\$50 point looks less like a competitive equilibrium than a jointly defended floor, held while both labs convert capability leads into revenue visibility for capital markets. The conditional discounts are how they keep large customers without touching the number an analyst will model.

VI. The escape valve is open, and it runs through Hangzhou

What stops a defended floor is a credible substitute, and one arrived four days after Astra. DeepSeek V4.1-Flash posts 74.2 on DeepSWE v1.1 — against 74.0 for Claude Opus 5 and 73.0 for GPT-5.6 Sol on the same benchmark — at \$0.60 per million output tokens off-peak. If that result survives independent replication on tasks that matter, the frontier premium is being charged for reliability, tooling and enterprise assurance rather than raw capability, and it is vulnerable.

Two forces push the other way. Regulation raises the cost of being a serious vendor: on 1 September the European Commission sent its first AI Act information requests to more than 30 providers, covering safety and copyright — compliance overhead that favours incumbents. And compute is contracted years ahead at scale; reporting this week identified Anthropic as the buyer in a \$13.7 billion six-year GPU agreement, a single-sourced claim that should be treated as unconfirmed. Fixed costs of that size are an argument for holding price, not cutting it.

FIG. 2: PRICE DECLINE BY TIER, MARCH 2023 TO 2026 (FACTOR OF REDUCTION)



BARS SCALED TO MAX (300x). "BUDGET" = CHEAPEST MODEL CAPABLE OF A GIVEN TASK; "FRONTIER" = BEST AVAILABLE MODEL, \$30 TO \$2.50 PER 1M. SOURCE: TOKENCOST AI PRICE INDEX, 2026.

VII. How it plays out: three scenarios

Probabilities below are analytical judgment, not measured likelihoods, on a horizon running to end-2028.

1 · Baseline — the floor holds ~55%

2026–2028 · STEADY

Frontier list price stays near \$10/\$50 through 2027 while discounting continues via cache, batch and off-peak levers. Enterprises respond by tiering aggressively and cost per workflow rises, though below Gartner's fivefold. Watch for: the November and December promotional expiries taking effect as scheduled; the next flagship from any of the three labs arriving at or above \$10/\$50; inference-tiering and model-routing tooling becoming a standard line item in enterprise AI stacks.

2 · Acceleration — the floor breaks ~20%

MID-2027 ONWARD · FAST

An open-weight or challenger model matches frontier agentic performance convincingly enough to force a cut, or post-IPO scale and new silicon make one affordable. Frontier list drops below \$6/\$30 unconditionally. Watch for: independent replication of DeepSeek-class agentic benchmarks on enterprise tasks; a lab publishing a permanent rather than promotional cut; hyperscalers guiding capex down while token volume keeps climbing.

3 · Stall — prices ratchet up ~25%

2027 · SLOW

Promotional rates lapse into real increases, long-context and latency surcharges spread across the tier, and rising cost per workflow collides with flat AI budgets. Enterprise deployment plateaus at the pilots that already clear the bar. Watch for: Gemini Flash and Sol repricing upward on schedule; surcharge structures appearing in rival flagships; enterprise AI spend guidance softening in Q1–Q2 2027 earnings.

WHAT WOULD FALSIFY THIS THESIS

The claim is narrow: *frontier-tier price competition has paused, and per-token deflation no longer predicts enterprise AI cost*. It is wrong if a top-three agentic model lists below \$5/\$25 unconditionally before mid-2027, or if Google and OpenAI extend their expiring promotional rates at the introductory level rather than letting them step up — either would show the floor was a coincidence of launch timing, not a defended position.

The strongest counter-case is already visible in the data and is not dismissed here: the frontier index is still 84% below its March 2023 baseline, and Astra's lower token consumption per task means cost per outcome can fall while list price sits still. If token-efficiency gains outrun workflow-complexity growth, the Gartner mechanism inverts and the premium tier gets cheaper in practice without any published cut. The two-week window this brief rests on is also short; a single quarter of flat pricing is a signal, not a trend.

WHAT TO WATCH, IN ORDER

1. **21 November 2026** — OpenAI's GPT-5.6 Sol price guarantee expires. Check the published API pricing page: does \$4/\$20 hold, or step up?
2. **31 December 2026** — Gemini 3.8 Flash introductory pricing ends and is scheduled to double to \$1.50/\$7.50. The first real test of whether cheap-tier prices are structural or promotional.
3. **October 2026** — Anthropic's Nasdaq listing, if it proceeds. The S-1 would be the first audited disclosure of frontier-model gross margin, settling whether \$10/\$50 is a floor or a ceiling.
4. **Late January 2027** — Q4 hyperscaler earnings and 2027 capex guidance, against the \$660–690bn 2026 base. Capex that keeps climbing is an argument for holding price.
5. **Next frontier release from any of the three labs** — the single cleanest indicator. A flagship priced below \$10/\$50 without conditions ends this thesis; one at or above it confirms the tier.

SOURCES

1. Anthropic, "Introducing Claude Fable 5.1 and Claude Mythos 5.1," 1 September 2026 · anthropic.com

2. LLM Stats, "GPT-6 Astra: benchmarks, pricing and context window," September 2026 · llm-stats.com
3. CloudZero, "GPT-6 Astra pricing: what OpenAI's new flagship costs in 2026," September 2026 · cloudzero.com
4. Yahoo Finance, "GPT-6 Astra pricing confirms OpenAI's premium track: \$10/\$50 while rivals cut," September 2026 · finance.yahoo.com
5. eesel AI, "GPT-6 Astra pricing: every API tier and ChatGPT plan in 2026," September 2026 · eesel.ai
6. GCN, "Google ships Gemini 3.8 Flash with stronger agentic coding at \$0.75 per million tokens," 2 September 2026 · gcn.com
7. Dataconomy, "DeepSeek launches V4.1-Flash with 1M-token context," 11 September 2026 · dataconomy.com
8. DataNorth, "DeepSeek-V4.1-Flash: MIT open weights at \$0.15 per million," September 2026 · datanorth.ai
9. BenchLM, "LLM API pricing trends and updates," 11 September 2026 · benchlm.ai
10. TokenCost, "AI Price Index: LLM costs 2023–2026," 2026 · tokencost.app
11. Gartner, "Gartner predicts AI inference costs per agentic workflow will increase more than fivefold through 2028," 17 August 2026 · gartner.com
12. Futurum Group, "AI capex 2026: the \$690B infrastructure sprint," 12 February 2026 · futurumgroup.com
13. Bloomberg, "SoftBank gets upsized \$11.9 billion loan in OpenAI funding push," 14 September 2026 · bloomberg.com
14. Business Today, "Anthropic sets sights on Nasdaq for potential October IPO, eyes \$2 trillion valuation," 14 September 2026 · businesstoday.in
15. Anthropic, "Anthropic invests \$50 billion in American AI infrastructure," 12 November 2025 · anthropic.com
16. Agence Europe, "European Commission sends first requests for information to more than 30 AI providers," 1 September 2026 · agenceurope.eu
17. AI Weekly, "AI news today, 14 September" (reporting The Information on Anthropic's \$13.7bn GPU agreement, 13 September 2026) · aiweekly.co

METHOD NOTE: FIGURES ARE ATTRIBUTED AND DATED INLINE; WHERE INDEPENDENT SOURCES DISAGREE, THE DISAGREEMENT IS REPORTED RATHER THAN AVERAGED. GPT-6 ASTRA'S RELEASE IS GIVEN AS 3 SEPTEMBER BY ONE TRACKER AND 4 SEPTEMBER BY ANOTHER; BOTH ARE CITED. THE \$13.7BN COMPUTE AGREEMENT RESTS ON SINGLE-SOURCE REPORTING AND IS FLAGGED AS UNCONFIRMED. SCENARIO PROBABILITIES ARE ANALYTICAL JUDGMENT.