LLM API Pricing 2026: A Reproducible Cost-per-Task Comparison

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · llm api pricing · cost per token · cost per task · openai pricing · anthropic pricing · gemini api pricing · llm inference cost · total cost of ownership · api pricing comparison

Executive Summary

As of September 2026, API pricing for large language models (LLMs) has fragmented into a wide band rather than converging on a single rate. Flagship reasoning models such as **OpenAI's GPT-6 Astra** list at $10 per million input tokens and $50 per million output tokens (developers.openai.com), while **Anthropic's Claude Opus 5** holds at $5 input / $25 output per million tokens, the same rate as its predecessor (qz.com). **Google's Gemini 3.1 Pro Preview** charges $2 input / $12 output per million tokens below a 200,000-token threshold, rising to $4/$18 above it (ai.google.dev). Mid-tier and economy models cluster far lower, from **GPT-5.6 Terra** and **Claude Sonnet 5** near $2/$10-12 down to **GPT-5.6 Luna** at $0.20/$1.20, and open-weight models hosted on Together AI run as low as $0.14/$0.28 for DeepSeek V4 Flash (together.ai).

Sticker price per token is a poor proxy for cost per completed task. Prompt caching, batch discounts, and hidden reasoning tokens can materially change effective cost. OpenAI, Anthropic, and Google all confirm that internal "thinking" or reasoning tokens are billed as output tokens even though they are invisible in the response

(developers.openai.com) (ai.google.dev), while batch APIs cut standard rates by roughly half across OpenAI, Anthropic, Google, Mistral, and Together AI (mistral.ai). Independent benchmark provider **Artificial Analysis** defines cost per task as a weighted-average figure that explicitly rises when a model produces more reasoning tokens at an identical per-token price (artificialanalysis.ai), and **Epoch AI** finds the price needed to hit a fixed capability milestone has fallen between 9x and 900x per year depending on the task (epoch.ai).

This report publishes a reproducible method for computing cost per completed task from a provider's published token rates, a stated workload, and disclosed caching and retry assumptions, and applies it across nine current model families from six providers plus a Together AI-hosted open-weight tier. Enterprise adoption data from **Menlo Ventures'** survey of 495 U.S. decision-makers shows generative AI spending rose 3.2x year-over-year to $37 billion in 2025, with the top three vendors now capturing 88% of enterprise LLM API usage (menlovc.com) (menlovc.com). The remainder of the report walks through each provider's current rate card, the modifiers that separate list price from delivered cost, the published methodology for computing cost per task, and the considerations that determine whether a commercial API or a self-hosted open-weight deployment is the cheaper option at production volume.

Introduction and Background

Every major LLM provider now publishes per-token pricing for its API, but the headline number on a pricing page answers a narrower question than most buyers are actually asking. A developer choosing between **OpenAI**, **Anthropic**, **Google**, **xAI**, and **Mistral** for a production workload needs to know the cost of a *completed task*, a support ticket answered, a document summarized, a code change reviewed, not the cost of an abstract million tokens. Those two numbers diverge sharply once reasoning tokens, prompt caching, retries, and batch processing enter the calculation, and providers do not disclose task-level costs directly; they disclose token rates and billing rules, and leave the arithmetic to the developer.

This report is a companion to, not a replacement for, IntuitionLabs' earlier pricing comparison of the Grok, Gemini, OpenAI, and Claude APIs, last updated February 28, 2026 (intuitionlabs.ai). Rather than restate that earlier snapshot, this report focuses on what has changed since (new model generations, OpenAI's July 2026 price cuts, Google's introductory Flash pricing) and, more importantly, on a reproducible method: what to measure, on what inputs, with what assumptions, so a reader can run the same calculation against next month's rate card rather than trusting a static table.

As of September 5, 2026 (the publication date of this report), pricing across the industry is best described as a multi-tier structure repeated by every vendor: a flagship reasoning-capable model priced from $2 to $10 per million input tokens, a mid-tier model an order of magnitude cheaper, and an economy or "nano" tier priced in cents per million tokens. Layered on top of that base structure are caching discounts of up to 90%, batch discounts of roughly 50%, and reasoning-token billing that can multiply the effective output cost of a single query several times over. Distinguishing a vendor's published list price from its effective delivered cost, and distinguishing a vendor's own claim from independently observed evidence, is the organizing task of the sections that follow.

OpenAI API Pricing: GPT-6 Astra and the GPT-5.6 Family

OpenAI's current flagship model, **GPT-6 Astra**, is priced at $10.00 per million input tokens and $50.00 per million output tokens for prompts up to 272,000 tokens, with cached input billed at $1.00 and cache writes at $12.50; prompts longer than that threshold are billed at double the input and cache rate and 1.5 times the output rate for the entire request (developers.openai.com) (developers.openai.com). The model carries a 1,050,000-token context window, a 128,000-token maximum output, and an April 30, 2026 knowledge cutoff, and cache-write tokens generally are billed at 1.25 times the uncached input rate (developers.openai.com) (developers.openai.com).

Below the flagship, OpenAI's **GPT-5.6** family runs three tiers: **Sol**, currently on promotional pricing of $4.00/$20.00 per million input/output tokens through at least November 21, 2026; **Terra**, at $2.00/$12.00; and **Luna**, the economy tier, at $0.20/$1.20 (developers.openai.com). These are not static: business broadcaster **CNBC** reported that on July 30, 2026, OpenAI cut Terra's price by 20% to its current $2/$12 rate and cut Luna's price by 80% to its current $0.20/$1.20 rate, with OpenAI stating its strategy remains "focused on advancing both capability and efficiency" generation over generation (cnbc.com) (cnbc.com).

OpenAI's Batch and Flex processing tiers are priced at 50% of standard rates, while Fast mode (renamed from Priority processing on July 30, 2026) doubles the applicable rate for lower latency and is unavailable for GPT-6 Astra when EU data residency is selected; regional processing endpoints that guarantee data residency separately carry a 10% price uplift for models released on or after March 5, 2026 (developers.openai.com). Critically for cost-per-task calculations, OpenAI's own documentation states plainly that reasoning tokens generated internally by GPT-6 Astra and the GPT-5.6 line are billed as output tokens even though they never appear in the visible API response, and developers must inspect the `output_tokens_details` field of the usage object to see how many were actually billed (developers.openai.com) (developers.openai.com). OpenAI recommends the `max_output_tokens` parameter as the practical lever for capping combined reasoning-plus-visible-output spend on a runaway completion (developers.openai.com).

Anthropic Claude API Pricing

Anthropic's current lineup, as of September 2026, comprises **Claude Opus 5** at $5.00 input / $25.00 output per million tokens ($0.50 cached input, $6.25 cache write), **Claude Sonnet 5** at $2.00/$10.00 ($0.20 cached, $2.50 cache write), and **Claude Haiku 4.5** at $1.00/$5.00 ($0.10 cached, $1.25 cache write) (platform.claude.com). Independent business outlet **Quartz** reported that Opus 5 launched at the same per-token price as its predecessor, Opus 4.8, while narrowing the capability gap to Anthropic's larger Fable 5 model at roughly half the price (qz.com). Anthropic's prompt-caching structure has more granularity than most rivals: a 5-minute cache write costs 1.25 times the base input price, a 1-hour cache write costs 2 times the base input price, and a cache read (hit) costs just 0.1 times the base input price on most models (platform.claude.com). Anthropic's Batch API applies a flat 50% discount on both input and output tokens; Batch-priced Opus 5, for example, runs $2.50/$12.50 per million tokens (platform.claude.com). Claude 4.6-and-later models, including Opus 5, bill the full 1-million-token context window at the standard per-token rate with no long-context surcharge, a structural difference from OpenAI's and Google's tiered long-context pricing. US-only data-residency inference carries a 1.1x multiplier on both input and output tokens, and a Fast mode for Opus 5 runs up to 2.5 times faster at double the base per-token price, corroborated independently by Quartz's reporting on the same launch (claude.com) (qz.com).

Extended thinking (Anthropic's term for reasoning tokens) follows the same billing logic as OpenAI's: thinking tokens count toward the `max_tokens` limit for the turn and are reported as billed output tokens, and Anthropic directs developers to the `usage.output_tokens_details.thinking_tokens` field to audit exactly how much of a bill was internal reasoning rather than visible answer text (platform.claude.com) (platform.claude.com).

Google Gemini API Pricing

Google prices its current flagship, **Gemini 3.1 Pro Preview**, at $2.00 input / $12.00 output per million tokens for prompts up to 200,000 tokens, rising to $4.00/$18.00 above that threshold, with context caching at $0.20/$0.40 per million tokens plus a separate $4.50-per-million-tokens-per-hour storage fee (ai.google.dev). The same rate appears on Google's enterprise pricing surface, the **Gemini Enterprise Agent Platform**, which is where Google's former Vertex AI generative-AI pricing page now redirects (cloud.google.com). Google's prior flagship, **Gemini 2.5 Pro**, remains listed at $1.25/$2.50 input and $10.00/$15.00 output across the same 200,000-token threshold, illustrating how the newest flagship generation can cost more per token even as capability improves (ai.google.dev).

Mid-tier **Gemini 3.5 Flash** is priced at $1.50 input / $9.00 output per million tokens, and the economy **Gemini 3.5 Flash-Lite** at $0.30/$2.50, with a slightly older **Gemini 3.1 Flash-Lite** available at $0.25/$1.50 for text, image, and video (ai.google.dev). Google's Batch API applies a 50% cost reduction across the paid tier, so Flash's batch price runs exactly $0.75/$4.50 against its standard $1.50/$9.00, and Google is separately running introductory pricing on its newer Flash releases at $0.75/$3.75 per million input/output tokens through December 31, 2026, doubling to $1.50/$7.50 on January 1, 2027 (ai.google.dev). Tech outlet **VentureBeat** covered the launch directly, confirming the $0.75/$3.75 introductory window and reporting that Google's own competitive-benchmark table places Claude Sonnet 5 at $2/$10 and GPT-5.6 Terra at $2/$12, framing the Flash cut as a direct bid for cost-sensitive coding and agent workloads (venturebeat.com) (venturebeat.com). As with OpenAI and Anthropic, Google's documentation is explicit that when thinking is enabled, response pricing sums output tokens and thinking tokens, and billing is based on the full internal thought token count rather than the shorter summary a user actually sees (ai.google.dev) (ai.google.dev).

Additional Providers: xAI, Mistral, and Open-Weight Model Hosting

xAI's flagship **Grok 4.6** is priced at $2.00 input / $6.00 output per million tokens below 200,000 tokens of context, rising to $4.00/$12.00 above that threshold, with cached input at $0.50 ($1.00 in the long-context tier); xAI's own API FAQ restates the headline $2/$6 figure directly (docs.x.ai) (x.ai). The prior-generation **Grok 4.3** and the **Grok 4.20** variants are priced at $1.25 input / $2.50 output, with a batch discount of 20% for those specific models; xAI states plainly that models not on the published batch-discount list receive no discount at all (docs.x.ai).

Mistral prices its **Mistral Large** model at $0.5 input / $1.5 output per million tokens by its own published example, and its current frontier-class model, **Mistral Medium 3.5**, at $1.5/$7.5 (mistral.ai) (docs.mistral.ai). Its economy-tier **Mistral Small 4**, a 119-billion-parameter hybrid instruct and reasoning model, runs $0.15/$0.6 (docs.mistral.ai). Mistral offers a 50% batch-processing discount and states that cached input tokens can reduce input cost by up to 90% for repeated prompts (mistral.ai) (mistral.ai).

Beyond the model developers themselves, inference hosts such as **Together AI** serve open-weight models at serverless rates that vary by model; those rates are not universally below frontier commercial APIs. Together AI's model library lists **DeepSeek V4 Flash** at $0.14 input ($0.03 cached) / $0.28 output per million tokens, **DeepSeek V4 Pro** at $1.32/$3.96, and Qwen3.8-2.4T-A95B at $2.50 input ($0.50 cached) / $6.25 output per million tokens (Together AI model library). Together AI's batch inference product advertises savings of up to 50% against its own real-time API for most serverless models, matching the batch economics offered by the frontier labs themselves (together.ai).

Comparative Context and Market Positioning

Table 1 below places each provider's current flagship reasoning model side by side on the dimensions that most affect cost per task: base per-token price, long-context surcharge behavior, and the caching and batch mechanisms available to lower effective spend.

Model	Input / 1M	Output / 1M	Long-context surcharge	Batch discount
GPT-6 Astra (OpenAI)	$10.00	$50.00	2x input, 1.5x output above 272K tokens	50%
Claude Opus 5 (Anthropic)	$5.00	$25.00	None (full 1M window at standard rate)	50%
Gemini 3.1 Pro Preview (Google)	$2.00	$12.00	2x input, 1.5x output above 200K tokens	50%
Grok 4.6 (xAI)	$2.00	$6.00	2x input, 2x output above 200K tokens	No batch discount
Mistral Medium 3.5 (Mistral)	$1.50	$7.50	Not published	50%

Figures in Table 1 are drawn from and cited to the official pricing pages discussed in the provider sections above. Flagship prices span a five-fold range on input tokens and a roughly eight-fold range on output tokens, with GPT-6 Astra the most expensive per token and Grok 4.6 and Gemini 3.1 Pro Preview the cheapest among reasoning-capable flagships. Claude Opus 5 stands out structurally for charging no long-context surcharge at all (platform.claude.com), which matters disproportionately for retrieval-augmented or document-heavy workloads that regularly exceed 200,000 tokens of context.

Table 2 (September 2026 rate card) extends the comparison to mid-tier, economy, and open-weight options, where the spread widens further.

Model	Input / 1M	Output / 1M	Tier
GPT-5.6 Terra (OpenAI)	$2.00	$12.00	Mid
GPT-5.6 Luna (OpenAI)	$0.20	$1.20	Economy
Claude Sonnet 5 (Anthropic)	$2.00	$10.00	Mid
Claude Haiku 4.5 (Anthropic)	$1.00	$5.00	Economy
Gemini 3.5 Flash (Google)	$1.50	$9.00	Mid
Gemini 3.5 Flash-Lite (Google)	$0.30	$2.50	Economy

Model	Input / 1M	Output / 1M	Tier
Mistral Small 4 (Mistral)	$0.15	$0.60	Economy
DeepSeek V4 Flash (open-weight, via Together AI)	$0.14	$0.28	Open-weight
Qwen3.8-2.4T-A95B (open-weight, via Together AI)	$2.50	$6.25	Open-weight

The economy tier is where per-token price becomes nearly incidental to the vendor's brand: OpenAI's Luna, at $0.20/$1.20, and the open-weight DeepSeek V4 Flash, at $0.14/$0.28, sit within a single order of magnitude of each other despite coming from a frontier lab and a specialist inference host respectively. That convergence at the bottom of the market, alongside real separation at the flagship tier, is consistent with Menlo Ventures' finding that enterprise buyers now concentrate spend on a small number of vendors for flagship work while treating economy-tier models as increasingly interchangeable commodities (menlovc.com).

Data Analysis and Evidence

Computing cost per completed task, rather than cost per token, requires a stated method. **Artificial Analysis**, an independent benchmark provider, defines its published Cost per Task metric as "the weighted-average cost (USD) to complete one Artificial Analysis Intelligence Index task," and states explicitly that models producing longer answers or more reasoning tokens will show a higher cost per task even when their per-token price is identical to a terser competitor's (artificialanalysis.ai) (artificialanalysis.ai). Stanford's **HAI AI Index Report 2025**, built on pricing data from Artificial Analysis and Epoch AI, uses a related but distinct method for tracking price over time: a 3:1 weighted average of input and output token prices at a fixed capability threshold, which found the price to query a model matching GPT-3.5-era MMLU performance fell more than 280-fold, from $20 per million tokens in November 2022 to $0.07 by October 2024 (hai.stanford.edu) (hai.stanford.edu). **Epoch AI's** own inference-price data insight finds that decline rate is highly task-dependent: the price to match GPT-4's performance on a set of PhD-level science questions fell roughly 40x per year, while across all milestones tracked the rate ranged from 9x to 900x per year, a spread wide enough that no single "price is falling by X% per year" headline is defensible without naming the specific task (epoch.ai) (epoch.ai).

A reproducible cost-per-task calculation, built from the rate cards above rather than an external benchmark score, has four inputs that must be disclosed before the number means anything: (1) input tokens per task, (2) output tokens per task including hidden reasoning tokens, (3) the cache-hit rate across repeated calls, and (4) the retry rate for calls that fail validation and must be resent. As a worked illustration **(Hypothetical Example)**, assume a document-summarization task that sends 4,000 input tokens and receives 800 output tokens per call, cached at a 70% hit rate on the input side, with a 10% retry rate. At Claude Sonnet 5's published rates ($2.00 input, $0.20 cached input, $10.00 output per million tokens), the blended cost per successful task is approximately: (0.3 x 4,000 x $2.00/1,000,000) + (0.7 x 4,000 x $0.20/1,000,000) + (800 x $10.00/1,000,000), multiplied by 1.10 to account for the retry rate, which comes to $0.012056 per completed summary. Running the identical assumptions through GPT-6 Astra's rate card ($10.00 input, $1.00 cached, $50.00 output) yields $0.06028 per task, exactly five times as high, driven mostly by Astra's output price rather than its input price. This is arithmetic on the publicly cited rate cards above, not a third-party benchmark result, and a reader can substitute their own measured token counts, cache-hit rate, and retry rate to reproduce or correct it for their own workload.

Reasoning tokens complicate this further because they are invisible in the response yet billed as output. Anthropic's documentation directs developers to the `usage.output_tokens_details.thinking_tokens` field specifically so that a reasoning-heavy call's true cost can be separated from its visible answer length (platform.claude.com). A separate framework, published on **arXiv**, models total cost of ownership for self-hosted inference from first principles, estimating hourly GPU cost as depreciation plus power consumption plus maintenance and proposing a neutral baseline of $0.79 per GPU-hour so that comparisons across different hardware and model choices are not distorted by one team's discounted cloud contract (arxiv.org) (arxiv.org). On the demand side, AI safety research organization **METR** measures a related but different quantity, "time horizon," defined as the length of task, by human completion time, that a model can complete with a 50% success rate; it is a capability metric rather than a cost metric, but it is frequently cited alongside price data because it lets a buyer ask whether a cheaper model can still clear the task-length bar a workload requires (metr.org).

Table 3 summarizes the cost-modifying mechanisms available from each provider, which is the piece of the calculation most often left out of simple per-token comparisons.

Provider	Cache discount	Batch discount	Reasoning tokens billed as output	Source
OpenAI	Cached input ~90% off list	50% (Batch/Flex)	Yes, confirmed in docs	(developers.openai.com)
Anthropic	90% off (cache read = 0.1x)	50%	Yes, confirmed in docs	(platform.claude.com)
Google	90% off (varies by model)	50%	Yes, confirmed in docs	(ai.google.dev)
xAI	~75% off on Grok 4.6	20% (select models only)	Yes—reasoning tokens are billed (listed separately from completion tokens)	(docs.x.ai)
Mistral	Up to 90% off	50%	Not separately documented	(mistral.ai)
Together AI (open-weight)	Model-dependent, ~80-90% off	Up to 50%	Not separately documented	(together.ai)

At xAI, models not listed in the published batch-discount table have no batch discount. On the enterprise demand side, Menlo Ventures' survey of 495 U.S. enterprise AI decision-makers, fielded November 7 to 25, 2025 in partnership with an independent research firm, found total enterprise generative AI spending grew 3.2x year-over-year to $37 billion in 2025, up from $11.5 billion in 2024, with Anthropic's share of enterprise LLM spend rising to 40% from 24% the prior year and 12% in 2023 (menlovc.com) (menlovc.com).

Implications and Future Directions

Two structural trends look likely to continue through 2027. First, promotional or introductory pricing is becoming a standard launch tactic rather than an exception: Google has published a specific expiration date for its introductory Flash pricing (discussed above), and OpenAI's GPT-5.6 Sol promotion is similarly guaranteed only through a stated date, meaning any cost-per-task figure computed from current rates should be treated as time-bound rather than

durable. Second, the gap between frontier commercial APIs and open-weight models hosted by specialist inference providers remains wide at the top of the market and nearly closed at the bottom: Epoch AI's data shows this compression is uneven across capability levels rather than a uniform industry-wide decline (epoch.ai).

That compression is also reshaping the build-versus-buy decision for organizations running LLM workloads at real volume. IntuitionLabs, a life-sciences and AI consultancy that advises regulated research organizations on AI infrastructure rather than selling a competing API itself, frames the tradeoff plainly: "Frontier APIs are priced for occasional use," and per-token cost stops being a rounding error once an organization moves from occasional queries to screening literature or processing entire document corpora (intuitionlabs.ai). Its own published comparison of open-weight serverless rates, verified against live provider pricing on August 12, 2026, illustrates the same pattern shown in Table 2 above: "an open-weight model of comparable capability typically runs a fraction of the per-token price of a frontier API," though the firm is careful to note that a hosted open-weight deployment is not architecturally identical to a fully self-hosted one, since prompts still travel to a third-party inference provider under contractual no-retention terms rather than staying entirely inside an organization's own network (intuitionlabs.ai) (intuitionlabs.ai). For teams evaluating that choice, the practical implication is that per-token price should be weighed against integration cost, data-residency requirements, and the volume threshold at which a cheaper token rate actually offsets the engineering effort of running a private deployment, rather than treated as a standalone deciding factor.

Looking forward, reasoning-token billing is likely to remain the single largest source of divergence between sticker price and delivered cost, since every major provider now bills invisible internal reasoning as output tokens, and none has published a discount specific to reasoning-token volume the way each has for caching and batch processing. Buyers who standardize on `max_output_tokens` caps, aggressive prompt-caching design, and batch processing wherever latency permits will likely see costs fall faster than list prices do, simply because those three levers compound.

Frequently Asked Questions (FAQs)

What is the cheapest LLM API for production use as of September 2026? Among frontier labs, OpenAI's GPT-5.6 Luna, at $0.20 input / $1.20 output per million tokens (see Table 2), is the least expensive current economy-tier model from a major lab. Open-weight models hosted on Together AI can run cheaper still, with DeepSeek V4 Flash at $0.14/$0.28 per million tokens (see Table 2). Which is actually cheapest for a given workload depends on cache-hit rate, retry rate, and whether the task needs the flagship model's reasoning quality at all.

How can LLM API costs be benchmarked? Follow the four-input method in the Data Analysis section above: measure actual input tokens, actual output tokens (including reasoning tokens, visible via each provider's usage object), the achievable cache-hit rate for the specific prompt pattern, and the retry rate for malformed or rejected responses, then apply the provider's published per-token rates including any caching and batch discounts. Artificial Analysis publishes a comparable metric across many models for buyers who want an external reference point rather than building the calculation from scratch (artificialanalysis.ai).

Is there a reliable LLM inference cost calculator? No single calculator is authoritative because reasoning-token counts and cache-hit rates are workload-specific; the reproducible method above, applied to a provider's official pricing page, functions as a calculator when populated with a team's own measured token counts.

What is the total cost of ownership for a self-hosted LLM API versus a commercial one? Total cost of ownership for self-hosting includes GPU depreciation, power, and maintenance, which one published framework estimates at a neutral baseline of $0.79 per GPU-hour for comparability across configurations, plus the platform engineering, identity integration, and support functions that a commercial API bundles into its per-token price (arxiv.org).

How does OpenAI GPT-6 Astra pricing compare with Claude Opus 5 and Gemini 3.1 Pro Preview? GPT-6 Astra is the most expensive of the three at $10/$50 per million input/output tokens, Claude Opus 5 is roughly half that at $5/$25, and Gemini 3.1 Pro Preview is the least expensive at $2/$12 below its 200,000-token threshold, as detailed in Table 1 above.

Conclusion

Cost-per-task comparisons that stop at a provider's headline per-token price systematically understate the true dispersion across the market, because caching, batch processing, reasoning-token billing, and long-context surcharges each move effective cost by a large and provider-specific factor before a single completed task is counted. Flagship models from OpenAI, Anthropic, and Google span roughly a five-fold range on input price and an eight-fold range on output price, while economy and open-weight tiers converge to within a single order of magnitude of each other regardless of vendor. The reproducible method set out in this report, disclosed input and output token counts, a stated cache-hit rate, and a stated retry rate, applied against a provider's published rate card, gives a buyer a number that can be checked and recomputed rather than taken on faith. As promotional pricing windows close and new model generations launch through the remainder of 2026 and into 2027, that discipline matters more than memorizing any single table of prices, because the table in this report, like the one it succeeds, will itself need updating within months.
