LLM Cost Per Million Tokens: 2026 Inference Pricing Benchmark

Published July 10, 2026 37 min read

Executive Summary

As of July 2026, the cost of large language model (LLM) inference spans nearly four orders of magnitude depending on model tier and deployment method. On the commodity end, **DeepSeek's** official API prices input tokens at $0.14 per million on a cache miss and just $0.0028 per million on a cache hit, with output tokens at $0.28 per million (Source: api-docs.deepseek.com. At the frontier end, **OpenAI's** flagship reasoning tier, gpt-5.5-pro, costs $30.00 per million input tokens and $180.00 per million output tokens (Source: developers.openai.com). In between sit workhorse models such as **Anthropic's Claude Sonnet 5**, priced at an introductory $2 per million input tokens and $10 per million output tokens through August 31, 2026, reverting to $3/$15 afterward (Source: www.anthropic.com), and **Google's Gemini 2.5 Pro**, billed at $2.25 per million input tokens and $18.00 per million output tokens for prompts under 200,000 tokens (Source: ai.google.dev).

This price dispersion is not static; it is the product of one of the steepest cost-decline curves documented in computing history. Andreessen Horowitz's "LLMflation" analysis found that a model matching GPT-3's November 2021 performance (MMLU score of 42, priced then at $60 per million tokens) cost just $0.06 per million tokens by late 2024 via Together.ai's hosted Llama 3.2 3B, a **1,000-fold decline in three years** (Source: a16z.com). Epoch AI's independent benchmark-based analysis corroborates the trend but finds it uneven: the price to match GPT-4's score on PhD-level science questions fell roughly 40x per year, with the full range across six benchmarks spanning 9x to 900x annually and a median of 50x per year (Source: epoch.ai). A 2026 arXiv analysis of 3,237 models proposes a "Tiered Super-Moore's Law," finding an approximately 600-fold aggregate price decline since 2020, with economy-tier models exhibiting a price half-life of 1.10 years while flagship reasoning models resist the trend because of a reasoning premium averaging 31.5 times non-reasoning prices (Source: arxiv.org).

For teams weighing self-hosting against API consumption, GPU rental prices form the other half of the equation. As of mid-2026, on-demand **NVIDIA H100** GPUs rent for as little as $2.89 per hour on RunPod's PCIe tier (Source: www.runpod.io), $3.99 per hour on Lambda's SXM cluster tier (Source: lambda.ai), and roughly $6.16 per GPU-hour on CoreWeave's 8-GPU HGX H100 nodes (Source: coreweave.com). At production throughput, vLLM's late-2025 wide expert-parallelism benchmarks sustained 2,200 tokens per second per H200 GPU on DeepSeek-style mixture-of-experts models (Source: vllm.ai), a figure that converts self-hosted GPU rental into an effective per-token cost that frequently undercuts API pricing at high volume, though it requires absorbing engineering, orchestration, and utilization risk that a managed API abstracts away.

Real-world migrations illustrate both the upside and the friction. A published teaching scenario modeling a technical-documentation SaaS company moving 85 million monthly tokens off GPT-4 Turbo and onto a self-hosted Ollama and Llama 3.3 70B stack, explicitly disclosed by its publisher as an illustrative, non-client-derived example rather than an audited case, projects a bill falling from $4,200 to $109 per month, a 97 percent reduction, over a six-day migration (Hypothetical Example) (Source: academy.talki-app.fr). Enterprise deployment firm LLMDeploy reports a separate, named client, Nocodo LTD, cutting its monthly AI infrastructure cost from $10,000 to $2,000, an 80 percent reduction, by moving to a self-hosted large model (Source: llmdeploy.to). These vendor-published case studies should be read as illustrative rather than generalizable, since compliance requirements, GDPR obligations, and existing GPU access shaped each decision as much as raw unit economics, and at least one of the examples is explicitly a modeled teaching scenario rather than a real client engagement. The remainder of this report quantifies current per-million-token pricing across every major provider, breaks down GPU-hour costs for self-hosted deployment, and works through the volume thresholds at which each approach becomes economical.

Introduction and Background

The unit "cost per million tokens" has become the de facto currency of the large language model (LLM) industry, replacing per-request or per-character billing as the standard way vendors price API access to models such as **GPT**, **Claude**, **Gemini**, and open-weight alternatives like **Llama** and **DeepSeek**. A token is a sub-word unit of text, roughly three to four characters or three-quarters of an English word on average, and virtually every commercial LLM provider bills separately for input tokens (the prompt and context sent to the model) and output tokens (the text the model generates), with output tokens typically priced three to ten times higher than input tokens because generation is more computationally expensive than reading a prompt.

This report answers a query that has become central to any team building on LLMs: what does inference actually cost in mid-2026, whether accessed through a hosted API or run on rented GPUs. The question has grown more complicated, not less, as the market has matured. Where 2023 offered a handful of proprietary APIs at broadly similar prices, by July 2026 buyers choose among dozens of frontier and open-weight models, a fragmented GPU cloud market with an 8-fold spread between the cheapest and most expensive providers for identical silicon (Source: deploybase.ai), and inference optimization frameworks such as **vLLM** and **TensorRT-LLM** that materially change the economics of self-hosting relative to a managed API.

The stakes of getting this comparison right have risen alongside usage volume. A team processing 200 million tokens a month at GPT-4o's $2.50 input and $10.00 output per-million rate (Source: artificialanalysis.ai is looking at a materially different monthly bill than a team on a commodity open-weight model priced at $0.15 input and $0.60 output, as gpt-oss-120B is on Together AI (Source: www.together.ai. Decisions made at this scale ripple through gross margins for AI-native products and through infrastructure budgets for enterprises embedding LLMs into existing workflows.

This report draws on official pricing pages from every major model provider and GPU cloud vendor, third-party benchmark trackers including Artificial Analysis and Epoch AI, peer-reviewed and preprint economic analyses of the inference market, and five documented case studies of organizations that migrated between API and self-hosted deployment. Every price is anchored to its "as of" date because, as the data below demonstrates, these figures move quickly and in only one direction: down. The report addresses not only the headline question of price per million tokens, but the closely related questions of GPU rental economics, the price-performance tradeoffs among current models, and the volume thresholds at which self-hosting begins to outperform a managed API on pure cost grounds.

Methodology and Scope

This analysis compares list prices published on vendor pricing pages, cross-checked against the Artificial Analysis benchmark tracker, which aggregates pricing and performance data across more than 570 models as of mid-2026 (Source: artificialanalysis.ai).Three pricing categories are covered: first-party hosted APIs from model developers (**OpenAI**, **Anthropic**, **Google**, **DeepSeek**, **Mistral**); third-party inference platforms that host open-weight models (**Together AI**, **Groq**, **OpenRouter**, **AWS Bedrock**); and raw GPU rental markets for teams that self-host (**Lambda**, **RunPod**, **CoreWeave**, **AWS EC2**).

All prices are quoted in **US dollars per million tokens (MTok)** unless otherwise noted, split into input (prompt) and output (completion) rates where providers differentiate them. Where a provider offers prompt caching, a mechanism that discounts repeated context such as system prompts or long documents, that rate is noted separately, since Artificial Analysis's methodology treats cache-hit pricing as materially different from base input pricing and warns that Anthropic, Google, and OpenAI each structure cache economics differently (Source: artificialanalysis.ai). GPU rental prices are quoted as on-demand, per-GPU, per-hour rates, normalized where a vendor only lists multi-GPU node pricing.

Two important caveats apply throughout. First, list prices are not always the price a high-volume buyer pays; enterprise contracts, reserved capacity, and negotiated volume discounts can shift effective rates substantially, particularly in the GPU cloud market where CoreWeave and Lambda both offer reserved and spot tiers at 40 to 60 percent discounts to on-demand rates (Source: coreweave.com). Second, price alone does not capture value: a model's tokens-per-dollar figure means little without accounting for how many tokens it needs to complete a given task, how fast it generates them,

and how often its answers are usable without retries. Where available, this report notes intelligence and throughput benchmarks alongside price to give a fuller price-performance picture, consistent with Epoch AI's finding that cost trends measured per benchmark task can diverge meaningfully from trends measured in raw price per token (Source: epoch.ai).

Current API Token Pricing Across Major Providers

Table 1 below summarizes list pricing for the most widely used proprietary and open-weight models as of July 2026, drawn directly from each vendor's published pricing page.

PROVIDER	MODEL	INPUT $/MTOK	OUTPUT $/MTOK	NOTES
Anthropic	Claude Opus 4.8	$5.00	$25.00	Cache write $6.25, cache read $0.50 (Source: www.anthropic.com)
Anthropic	Claude Sonnet 5	$2.00 (intro) / $3.00 (standard)	$10.00 (intro) / $15.00 (standard)	Introductory rate expires August 31, 2026 (Source: www.anthropic.com)
Anthropic	Claude Haiku 4.5	$1.00	$5.00	Fastest, most cost-efficient current Claude tier (Source: www.anthropic.com)
OpenAI	gpt-5.6-terra	$2.50	$15.00	Mid-tier current-generation model (Source: developers.openai.com)
OpenAI	gpt-5.6-luna	$1.00	$6.00	Budget current-generation tier (Source: developers.openai.com)
OpenAI	gpt-5.5-pro	$30.00	$180.00	Frontier reasoning tier (Source: developers.openai.com)
OpenAI	GPT-4o (legacy)	$2.50	$10.00	Cache hit $1.50 (-40%) (Source: artificialanalysis.ai
Google	Gemini 2.5 Pro	$2.25 / $4.50	$18.00 / $27.00	Higher rate applies above 200k-token prompts (Source: ai.google.dev)
Google	Gemini 2.5 Flash	$0.54 (text)	$4.50	Audio input priced at $1.80 (Source: ai.google.dev)
DeepSeek	deepseek-chat (V3.2-class)	$0.14 (cache miss) / $0.0028 (cache hit)	$0.28	Official first-party pricing (Source: api-docs.deepseek.com)
Mistral	Mistral Large	$2.00	$6.00	Batch processing gets 50% discount (Source: mistral.ai)
Together AI	Llama 3.3 70B	$1.04	$1.04	Flat blended rate (Source: www.together.ai
Together AI	gpt-oss-120B	$0.15	$0.60	Open-weight OpenAI model (Source: www.together.ai
Groq	Llama 3.3 70B Versatile	$0.59	$0.79	Served at 394 tokens/sec (Source: groq.com)
OpenRouter	Llama 3.3 70B Instruct	$0.10	$0.32	131k context window (Source: openrouter.ai)
AWS Bedrock	Claude Opus 4.8	$6.00	$30.00	Bedrock carries a markup over Anthropic's direct API (Source: aws.amazon.com)
AWS Bedrock	DeepSeek V3.2	$0.62	$1.85	US-region Bedrock hosting (Source: aws.amazon.com)

The spread in Table 1 illustrates the central tension buyers face: the cheapest model in the table, DeepSeek's cache-hit input rate, is more than 10,000 times less expensive per token than OpenAI's most expensive frontier output rate. That gap is not solely a story about capability; it also reflects a two-tier market structure. Epoch AI's methodology explicitly excludes reasoning models from its per-token price comparisons because they generate far more tokens per answer, making raw per-token price misleading for cross-tier comparison (Source: epoch.ai). A buyer choosing between DeepSeek and GPT-5.5-pro is not just choosing a price point; they are choosing how many tokens the model will need to spend to reach a usable answer, and how much that answer is worth.

It is also worth noting the AWS Bedrock markup pattern visible in the table: Anthropic's Claude Opus 4.8 costs $6.00/$30.00 per million tokens through Bedrock versus $5.00/$25.00 directly through Anthropic's own API, a roughly 20 percent premium for the convenience of AWS billing, IAM integration, and regional data residency. Enterprises already committed to AWS often accept this premium in exchange for procurement simplicity, a pattern echoed in Azure OpenAI Service pricing, where the regional-deployment tier of GPT-4o costs $2.75 per million input tokens versus $2.50 for OpenAI's equivalent global-tier direct API pricing (Source: azure.microsoft.com).

A similar markup, and occasional pricing mismatch, shows up in Mistral's numbers. Mistral's own pricing page states plainly that "we charge per million tokens processed" and that Mistral Large costs $2 per million input tokens and $6 per million output tokens directly through Mistral's API (Source: mistral.ai, while AWS Bedrock lists a differently versioned "Mistral Large 3" at $0.50 input and $1.50 output per million tokens (Source: aws.amazon.com), a reminder that model version numbers and hosting-platform naming conventions do not always align cleanly across vendor and reseller pricing pages, and that buyers comparing quotes across platforms need to confirm they are pricing the identical model checkpoint rather than a same-named but distinct release. Google's own small-model lineup shows a similar tiering: its open-weight Gemma 3 27B model is priced on AWS Bedrock at $0.23 per million input tokens and $0.38 per million output tokens in Google's lowest-cost listed region (Source: aws.amazon.com), positioning it close to DeepSeek and the smallest Together AI and Groq open-weight offerings, while Google's proprietary Gemini 2.5 Flash and 2.5 Pro tiers remain priced well above that commodity floor as shown in Table 1. Together AI's own catalog illustrates a comparable spread among open-weight mixture-of-experts models: its hosted DeepSeek V4 Pro costs $1.74 per million input tokens, or $0.20 per million on a cache hit, and $3.48 per million output tokens (Source: www.together.ai, noticeably higher than DeepSeek's own first-party pricing for its chat-tier model, underscoring that third-party hosting of an open-weight model does not automatically match the price the model's own developer can offer.

GPU Cloud Costs for Self-Hosted Inference

Self-hosting shifts the cost question from "price per million tokens" to "price per GPU-hour," which then has to be converted into an effective token cost based on the throughput a given model and hardware combination can sustain. *Table 2* summarizes on-demand GPU pricing from three major GPU cloud vendors as of July 2026.

PROVIDER	GPU	ON-DEMAND $/HR	SPOT/DISCOUNTED $/HR
Lambda	H100 SXM (8-GPU cluster)	$3.99	Reserved clusters available on request (Source: lambda.ai)
Lambda	B200 SXM6	$6.69	n/a (Source: lambda.ai)
Lambda	A100 SXM (80GB)	$2.79	n/a (Source: lambda.ai)
RunPod	H100 SXM (Secure Cloud)	$2.99	Per-second billing available (Source: www.runpod.io
RunPod	H100 PCIe	$2.89	n/a (Source: www.runpod.io
RunPod	A100 PCIe (80GB)	$1.39	n/a (Source: www.runpod.io
RunPod	H200 SXM	$4.39	n/a (Source: www.runpod.io
CoreWeave	HGX H100 (8-GPU node, per GPU)	$6.16	$2.46 spot (Source: coreweave.com)
CoreWeave	HGX H200 (8-GPU node, per GPU)	$6.31	$2.62 spot (Source: coreweave.com)
CoreWeave	A100 (8-GPU node, per GPU)	$2.70	$1.21 spot (Source: coreweave.com)

Independent GPU cloud tracker DeployBase, which monitors 28 or more providers, put the March 2026 spread for a single H100 GPU-hour at $1.38 to $11.68, an 8.5-fold difference for identical NVIDIA Hopper-architecture silicon, attributing the gap to differences in managed services, SLAs, and how much surrounding infrastructure a provider bundles in (Source: deploybase.ai). A separate industry tracker cited by Thunder Compute placed AWS EC2's per-GPU H100 rate at approximately $6.88 to $12.29 per hour when normalized from AWS's 8-GPU P5 instances, roughly three times CoreWeave's node-normalized rate for the same chip (Source: www.thundercompute.com. Amazon's own EC2 P5 product page, powered by NVIDIA H100 and H200 GPUs, claims cost reductions of up to 40 percent on deep learning training relative to the prior generation but does not list a public on-demand hourly rate, requiring buyers to use AWS's pricing calculator or a reserved capacity block (Source: aws.amazon.com).

Reserved and spot pricing widen the effective gap further for teams able to commit to capacity in advance. CoreWeave's published rate card shows its on-demand HGX H100 node price of $49.24 per hour falling to a spot rate of $19.71 per hour, a 60 percent discount, and its A100 nodes falling from $21.60 to $9.65 per hour on the spot market (Source: coreweave.com). Independent tracker GPU Tracker, which monitors more than 54 providers and over 5,200 live instance listings, reported H100 pricing spanning from $0.80 to $97.44 per hour across its dataset, underscoring that headline on-demand rates from major branded providers represent only the middle of a much wider distribution that includes smaller marketplace resellers at both ends (Source: gputracker.dev. A separate market report focused specifically on H100 pricing during June 2026 put the median on-demand price across seven sourced provider rows at $4.29 per GPU-hour, with a public list-price band of $3.29 to $12.29 across five providers, explicitly cautioning that this median is "a sourced market indicator, not a transaction price," since individual buyers still negotiate on region, access mode, cluster size, support terms, and availability (Source: www.computetape.com).

Raw GPU-hour price is only half the calculation; the other half is throughput, meaning how many tokens per second a given model can generate on that hardware. vLLM's engineering team reported that its late-2025 wide expert-parallelism optimizations, including dual-batch overlap and DeepEP all-to-all kernels, lifted sustained throughput on a CoreWeave H200 cluster to 2,200 tokens per second per GPU for DeepSeek-style mixture-of-experts models, up from approximately 1,500 tokens per second per GPU in earlier benchmarks (Source: vllm.ai). At CoreWeave's $6.31 per-GPU-hour rate for H200 nodes, that throughput implies a raw compute cost on the order of $0.0008 per thousand output tokens before accounting for underutilization, batching overhead, or engineering cost, dramatically below any hosted API's output price, though only achievable at the batch sizes and utilization rates the benchmark assumes. A comparative academic study of open-source serving frameworks found that vLLM achieved up to 24 times higher throughput than Hugging Face's Text Generation Inference (TGI) under high-concurrency workloads, though TGI showed lower tail latency for single-user interactive scenarios, underscoring that self-hosted economics depend heavily on the serving stack chosen, not just the GPU rented (Source: arxiv.org).

Self-Hosted vs API: Break-Even Economics

The decision between consuming a hosted API and self-hosting on rented or owned GPUs hinges on volume, latency sensitivity, and how much engineering time a team can dedicate to inference optimization. At low volume, the fixed cost of provisioning and maintaining a GPU cluster, whether spot or reserved, dwarfs any per-token savings; a single idle H100 at RunPod's $2.89 hourly rate costs roughly $2,080 per month whether or not it processes a single token (Source: www.runpod.io. A team processing under a few million tokens per day rarely clears that fixed cost against even a mid-priced API such as Claude Sonnet 5 at $2/$10 per million (Source: www.anthropic.com).

The crossover point shifts as volume climbs into the hundreds of millions of tokens per month. A modeled teaching scenario published by an AI training academy, describing a fictional "TechDocs SaaS" technical-documentation generator and explicitly labeled by its publisher as an illustrative example rather than data from a real client (Hypothetical Example), projects the economics concretely: at 85 million tokens per month split roughly 60 million input and 25 million output, the modeled company's OpenAI GPT-4 Turbo bill reaches $4,200 per month before a migration to a self-hosted Ollama deployment running Llama 3.3 70B brings the same workload down to $109 per month, a 97 percent reduction over a modeled six-day engineering sprint with a projected 2.8-month return on investment (Source: academy.talki-app.fr). The same scenario models a 44 percent latency improvement after migration, since local inference eliminates the network round-trip to a third-party API, and projects output quality retained at 97 percent of the original GPT-4 Turbo baseline.

Not every workload benefits equally from migration. A separate account from a SaaS engineering team found that roughly 80 percent of its $4,200 monthly OpenAI spend was concentrated in a single low-complexity feature, internal documentation search, that was paying frontier-model output rates of $5 to $30 per million tokens for what was functionally extractive summarization, while the genuinely hard-reasoning 20 percent of traffic justified the frontier model's premium (Source: medium.com). This pattern, routing only the workloads that genuinely require frontier capability to the expensive API while shifting high-volume, lower-complexity tasks to cheaper or self-hosted models, has become a common cost-optimization strategy as the price gap between commodity and frontier models has widened rather than narrowed. A separate first-person account from a seed-stage startup's chief technology officer describes a comparable single-sprint migration off GPT-4o, which the post notes costs $2.50 per million input tokens and $10.00 per million output tokens, that the author claims cut the company's total LLM bill by a factor of 40 once structured-extraction workloads were moved to cheaper models and only genuinely hard tasks were retained on the frontier API (Source: dev.to).

The comparison is also not purely financial. A blog account from Herodesk's developer described migrating away from OpenAI primarily to achieve full GDPR compliance rather than to cut costs, noting that OpenAI's per-token rates at the time ranged from $0.75 to $5 per million input tokens and $5 to $30 per million output tokens across its model tiers (Source: anderseiler.com), a range that remains broadly consistent with OpenAI's July 2026 published pricing. Enterprise-focused vendor LLMDeploy separately reported that a client, Nocodo LTD, cut its monthly infrastructure cost from $10,000 to $2,000, an 80 percent reduction, and observed 3 times faster API response times after moving from OpenAI's hosted API to a customized self-hosted deployment tuned for the client's specific workload (Source: llmdeploy.to). Because these case studies are self-published by the companies or vendors involved, their figures should be read as directional evidence of the scale of savings possible, not as independently audited benchmarks; none of the accounts publish raw request logs or third-party cost audits.

Analysis of Key Market Segments

The pricing data assembled above sorts naturally into three tiers that behave differently over time. The **commodity tier**, occupied by small open-weight models and aggressively priced first-party offerings like DeepSeek's cache-hit rate of $0.0028 per million input tokens (Source: api-docs.deepseek.com, is where the 2026 arXiv analysis found the fastest price decay, with an estimated price half-life of 1.10 years, meaning commodity-tier prices roughly halve every 13 months (Source: arxiv.org). The **mid-tier or workhorse segment**, populated by models like Claude Sonnet 5, Gemini 2.5 Flash, and Llama 3.3 70B, showed a slower 1.55-year half-life in the same analysis, reflecting a market where vendors still capture some margin on differentiated quality (Source: arxiv.org).

The **frontier reasoning tier**, represented by OpenAI's gpt-5.5-pro at $30/$180 per million tokens (Source: developers.openai.com) and Anthropic's higher-end offerings, behaves almost independently of the broader cost-decline trend. The same arXiv analysis found a near-zero exponential price-decline fit for flagship models, with an R-squared of just 0.031, attributing this to what the authors term a "reasoning premium" that averages 31.5 times the price of non-reasoning models with comparable parameter counts (Source: arxiv.org. In effect, vendors appear to be conceding the low end of the market to commoditization while defending price on the newest, highest-capability tier, a dynamic the a16z analysis noted as early as 2024 when it observed that OpenAI's o1 model carried the same per-output-token price as GPT-3 had at its 2021 launch, $60 per million (Source: a16z.com.

A fourth, cross-cutting segment worth isolating is the GPU cloud market feeding self-hosted inference, which has itself bifurcated. One analysis of the market described a split between "neoclouds," specialist GPU providers like Lambda, RunPod, and CoreWeave that undercut the hyperscalers by 3 to 6 times for identical hardware, and the traditional hyperscaler tier of AWS, Azure, and Google Cloud, which bundle GPU access with broader managed-service ecosystems at a substantial premium (Source: alatirok.com). That same source estimated the median on-demand H100 price across its tracked providers at approximately $2.95 per GPU-hour in May 2026, down from above $7 in early 2024, a roughly 58 percent decline in about two years, tracking closely with the broader GPU price-performance improvements documented industry-wide (Source: alatirok.com).

Another independent GPU price tracker, monitoring 21 providers across 277 tracked instances, reported the cheapest confirmed in-stock H100 rate as $1.33 per hour on the Vast marketplace as of July 2026, against a six-month price range of $2.00 to $15.20 per hour and a broader on-demand spread from $1.29 to $127.82 per hour once every tracked provider and configuration is included (Source: gpufinder.dev). The same tracker noted that Vast, PrimeIntellect, and Digital Ocean were the most reliably in-stock providers over the trailing 30 days, while some listed vendors were frequently waitlisted, a reminder that headline low prices on GPU marketplaces do not always translate into available, ready-to-use capacity (Source: gpufinder.dev). On the demand side of the same market, Groq's tokens-as-a-service pricing illustrates how far commodity open-weight pricing has fallen for smaller models specifically: its OpenAI-derived gpt-oss-20B model runs at 1,000 tokens per second for $0.075 input and $0.30 output per million tokens (Source: groq.com, while its Qwen3 32B offering runs at 662 tokens per second for $0.29 input and $0.59 output per million (Source: groq.com.

Data Analysis and Evidence

The magnitude of the LLM price decline is large enough that different methodologies, applied by different research groups to different datasets, converge on broadly the same order of magnitude even though their precise multipliers vary. Stanford's HAI-affiliated tracking, as summarized by third-party analysis, found that the cost of reaching GPT-3.5-class performance (an MMLU score of 64.8) fell from $20.00 per million tokens in November 2022 to $0.07 per million tokens by October 2024, a decline of roughly 280 times in under two years (Source: report-ai.org. Andreessen Horowitz's independent methodology, using MMLU threshold analysis on a separate dataset limited to OpenAI, Anthropic, and Meta models, arrived at a comparable headline: the cost of an MMLU-42-level model fell from $60 per million tokens for GPT-3 at its November 2021 launch to $0.06 per million tokens for Together.ai's hosted Llama 3.2 3B by late 2024, a 1,000-fold decline (Source: a16z.com). At a higher performance bar, MMLU score of 83 (achievable only since GPT-4's March 2023 release), a16z found prices fell by a factor of 62 over a shorter window, still consistent with the firm's headline claim of roughly a 10-fold price decline per year for equivalent-performance models (Source: a16z.com).

Epoch AI's benchmark-anchored methodology, which combines its own evaluation database with Artificial Analysis pricing data across 36 unique price-performance observations, found more variation by task: decline rates ranged from 9 times per year on some benchmarks to 900 times per year on others, with a median of 50 times per year across all six benchmarks tested (MMLU, GPQA Diamond, MATH-500, MATH level 5, HumanEval, and Chatbot Arena Elo) (Source: epoch.ai). Critically, Epoch found that the fastest declines, those above roughly 200 times per year, only began appearing in data after January 2024, and that restricting the analysis to post-2024 data alone roughly quadrupled the median decline rate from 50 times to 200 times per year, suggesting the price collapse has itself been accelerating rather than holding at a constant rate (Source: epoch.ai).

A separate 2025 arXiv study by researchers including MIT's Neil Thompson, drawing on the same Artificial Analysis and Epoch AI data sources, estimated the price for a fixed level of benchmark performance falls 5 to 10 times per year for frontier models across knowledge, reasoning, math, and software-engineering benchmarks, and, after controlling for hardware price declines by isolating open-weight models, attributed roughly a 3-fold-per-year improvement specifically to algorithmic efficiency gains rather than cheaper chips (Source: arxiv.org). This finding is corroborated by the 2026 "Tiered Super-Moore's Law" analysis, which performed a cost decomposition finding that total factor productivity residuals, essentially software and architectural innovation rather than raw silicon improvement, account for approximately 103.7 percent of observed cost reduction, with GPU hardware contributing a statistically negligible -0.9 percent to the decline (Source: arxiv.org). That paper also identifies May 2024 as a structural break point in the pricing data using a Chow test (F-statistic of 5.74, p-value of 0.005), marking a transition from a technology-driven to a competition-driven pricing regime as open-weight models from Meta, Mistral, and Chinese labs like DeepSeek began pressuring incumbent proprietary pricing (Source: arxiv.org.

A further independent tracker, aggregating figures it attributes to Stanford HAI's AI Index alongside a16z and Epoch AI, summarized the same overall pattern, noting that costs fall roughly ten times per year at a fixed capability bar while cautioning that "frontier prices fall slower" and that "the cliff is in capability-per-dollar, not the sticker price of the newest model" (Source: report-ai.org). A separate trajectory analysis pegged GPT-4-equivalent inference cost at approximately $60 per million input tokens in November 2021, falling to roughly $5 by the GPT-4o launch in May 2024, and to a range of $0.40 to $2.50 per million tokens by May 2026 depending on whether a commodity or mid-tier model is used to hit the same capability bar (Source: presenc.ai). DeepSeek's own historical pricing announcements illustrate how far even its own commodity-tier pricing has fallen since launch: when DeepSeek-V3 first shipped, the company set output pricing at $1.10 per million tokens with cache-miss input at $0.27 per million and cache-hit input at $0.07 per million, explicitly billing the launch pricing as "still the best value in the market" (Source: api-docs.deepseek.com), a rate that its current $0.28 per million output and $0.0028 per million cache-hit input pricing has since undercut further (Source: api-docs.deepseek.com).

On the compute side, GPU cloud pricing shows a parallel but more modest decline. GPU price tracker GetDeploying, monitoring 47 or more providers, reported average H100 pricing around $3.46 per hour with a low of $0.34 per hour on spot instances as of mid-2026, a spread the tracker attributes primarily to reliability and commitment-term differences rather than raw hardware cost (Source: getdeploying.com. A16z's own factor breakdown of the LLM cost decline attributes the trend to five distinct forces operating simultaneously: improved GPU cost-per-operation, model quantization (moving from 16-bit to increasingly 4-bit inference on newer hardware), software and memory-bandwidth optimizations, smaller and more efficiently trained models, and the emergence of open-weight competition that compresses margin across the value chain (Source: a16z.com).

Case Studies and Real-World Examples

TechDocs SaaS: A Modeled Migration from OpenAI to Self-Hosted Ollama (Hypothetical Example)

Talki Academy, an AI training publisher, presents "TechDocs SaaS," a technical-documentation generation platform, as a modeled teaching scenario, explicitly disclosed on the page itself as a "fictional case, teaching scenario" whose "figures below are illustrative and do not come from a real client" (Source: academy.talki-app.fr). The scenario models the company running its generation pipeline on OpenAI's GPT-4 Turbo, processing approximately 85 million tokens per month (60 million input, 25 million output) across 180,000 monthly requests, with a bill that climbs to $4,200 per month, representing a modeled 28 percent of revenue (Source: academy.talki-app.fr). Over a modeled six-day engineering sprint, the scenario walks through a migration to a self-hosted stack running Ollama with Llama 3.3 70B on dedicated GPU infrastructure, using a phased rollout with A/B testing against the existing GPT-4 Turbo pipeline. The projected result is a new monthly cost of $109, a 97 percent reduction, alongside a 44 percent latency improvement and output quality assessed at 97 percent of the original baseline, with a modeled payback period of 2.8 months on the migration engineering cost (Source: academy.talki-app.fr. Because the publisher itself labels the figures illustrative rather than empirical, this scenario is presented here as a worked cost model rather than as evidence of a specific company's actual outcome.

Nocodo LTD: Enterprise Infrastructure Consolidation with LLMDeploy

Nocodo LTD, an enterprise client of AI infrastructure vendor LLMDeploy, had grown its OpenAI API usage to a monthly cost of $10,000, a figure the company's case study describes as threatening its ability to scale AI-powered features cost-effectively, alongside secondary concerns about rate-limit predictability, data privacy for sensitive customer information, and compliance requirements tied to keeping data inside its own infrastructure (Source: llmdeploy.to). LLMDeploy's engagement involved an infrastructure assessment of Nocodo's existing server capacity followed by deployment of a customized open-source model tuned to the company's specific workload with load balancing and caching. The reported outcome was a reduction in monthly cost to $2,000, an 80 percent decrease, alongside a threefold improvement in API response time (Source: llmdeploy.to).

Lyzr.ai's NeoAnalyst: Compliance-Driven Migration to LLaMA2

Lyzr.ai, an Antler-backed enterprise generative AI company, built its NeoAnalyst natural-language data analytics platform on GPT-4, but found enterprise customers increasingly required an open, self-hosted deployment to satisfy GDPR data-privacy and SOC2 security requirements, constraints a proprietary hosted API structurally could not meet (Source: www.goml.io). The company's own account frames the migration to a self-hosted LLaMA2 deployment as driven jointly by compliance and cost, citing the GPT-4 API's high usage costs as placing pressure on margins alongside the closed architecture's limits on fine-tuning flexibility for orchestrating multiple AI agents (Source: www.goml.io). Lyzr reports the migration delivered a 30 percent cost reduction in the enterprise SaaS analytics deployment.

Herodesk: GDPR Compliance as the Primary Migration Driver (Hypothetical Example Excluded, Named Case)

Herodesk, a customer-support SaaS product, documented in a developer's public blog post moving its AI-powered features off OpenAI roughly two years after first integrating the OpenAI SDK for translation and classification tasks. The post frames the original OpenAI pricing, which it describes as ranging between $0.75 and $5 per million input tokens and $5 and $30 per million output tokens depending on model tier (Source: anderseiler.com, as reasonable in isolation but ultimately incompatible with the company's goal of full GDPR compliance for a fully self-hosted, EU-based data pipeline. The account is presented as a real, named migration rather than a hypothetical, illustrating that cost is frequently a secondary rather than primary driver behind self-hosting decisions.

Vendor Concentration Risk: A 680-Million-Token Dependency

A separate account describes an unnamed company processing 680 million tokens per month through OpenAI's API with no fallback provider, a dependency that became an acute concern when reports of internal cost-cutting pressure at OpenAI circulated publicly, prompting the company's leadership to conclude that switching providers if OpenAI's pricing or availability changed unfavorably would take six to eight months at minimum under its existing architecture (Source: michaeleakins.com). This example, while less about direct cost comparison than the others, illustrates a second economic dimension of the buy-versus-build decision: single-provider API dependency carries a switching-cost risk that does not appear on a monthly invoice but materially affects the total cost of ownership calculation, particularly for companies whose token volume has grown large enough that a provider's pricing or strategy shift would be difficult to absorb quickly.

Implications and Future Directions

The data assembled in this report point toward a market that is bifurcating rather than converging. Commodity and mid-tier inference is approaching a cost floor low enough that, per a16z's estimate, a full day of continuous human speech could be processed by a GPT-3-class model for about $2 per year, and the entire Linux kernel's source code could be processed for under $1 (Source: a16z.com). At that price level, entire categories of previously uneconomical LLM applications, high-frequency document processing, per-message classification, and always-on agentic monitoring, become viable at consumer or small-business budgets. Simultaneously, frontier reasoning models are moving in the opposite direction, with the 31.5-times reasoning premium documented in the 2026 arXiv analysis suggesting that vendors have found a segment of buyers willing to pay a substantial premium for the newest reasoning capability regardless of the commodity floor beneath it (Source: arxiv.org).

For infrastructure planning, this bifurcation argues for a routing architecture rather than a single-model commitment. Several of the case studies in this report describe exactly this pattern emerging organically: a company that discovers a large share of its spend goes to frontier-tier pricing for tasks that a commodity or self-hosted model can handle just as well, and subsequently builds a router that sends only genuinely hard reasoning tasks to the

expensive model. This mirrors the broader industry direction visible in OpenAI's own multi-tier pricing structure, which as of July 2026 spans from gpt-5.6-luna at $1.00/$6.00 to gpt-5.5-pro at $30.00/$180.00 within a single vendor's catalog, a 30-fold input-price spread that only makes sense if buyers are expected to route different task types to different tiers (Source: developers.openai.com).

On the GPU cloud side, the widening gap between hyperscaler and neocloud pricing, roughly 4.6 times for identical H100 silicon between Azure and RunPod per one industry analysis (Source: alatirok.com), suggests self-hosting economics will keep improving fastest for teams willing to work with specialist GPU providers rather than defaulting to the hyperscaler they already use for other infrastructure. Continued throughput gains from serving frameworks, illustrated by vLLM's move from roughly 1,500 to 2,200 tokens per second per H200 GPU in the space of months through kernel-level optimization alone (Source: vllm.ai), mean that the break-even volume at which self-hosting beats a hosted API is itself falling over time, independent of any change in GPU rental price. Teams evaluating the buy-versus-build decision in 2026 should treat both sides of the comparison as moving targets and revisit the calculation on a quarterly basis rather than treating it as a one-time architectural decision.

The Groq pricing table adds a further wrinkle worth flagging for buyers optimizing on latency as well as price: Groq's specialized inference hardware serves Llama 3.1 8B Instant at 840 tokens per second for $0.05 input and $0.08 output per million tokens, and Llama 3.3 70B Versatile at 394 tokens per second for $0.59/$0.79 per million (Source: groq.com), figures that place Groq's smallest model among the least expensive hosted options in this report while also delivering among the highest published throughput of any provider surveyed. For latency-sensitive applications such as conversational agents or real-time agentic tool-calling loops, this combination of low price and high tokens-per-second illustrates that the cheapest model per token and the cheapest model per completed interaction are not always the same choice, since a slower model may need more wall-clock time, and therefore more compute-hours, to finish an equivalent task even at a lower headline per-token rate.

Looking further ahead, the cost decomposition finding that software and algorithmic efficiency, not hardware, account for the overwhelming majority of the price decline (Source: arxiv.org) implies the decline is not obviously capped by physical hardware limits the way, for instance, Moore's Law eventually was. Whether the current pace persists depends on continued gains from quantization, mixture-of-experts routing, speculative decoding, and competitive pressure from open-weight models, all factors that could plausibly continue for several more years but that no forecaster, including Epoch AI's own methodology notes, claims to predict with confidence (Source: epoch.ai).

Frequently Asked Questions (FAQs)

What is the average cost per million tokens for LLM inference in 2026? There is no single average, because pricing spans roughly four orders of magnitude by design. Commodity models like DeepSeek's cache-hit input rate cost $0.0028 per million tokens (Source: api-docs.deepseek.com, workhorse models like Claude Sonnet 5 sit around $2 to $3 per million input tokens (Source: www.anthropic.com), and frontier reasoning models such as gpt-5.5-pro reach $30 per million input tokens and $180 per million output tokens (Source: developers.openai.com).

How does GPT-4o's cost per million tokens compare to newer models? GPT-4o is priced at $2.50 per million input tokens and $10.00 per million output tokens on OpenAI's direct API, according to Artificial Analysis's tracked pricing (Source: artificialanalysis.ai, placing it in the same price band as OpenAI's newer gpt-5.4 tier at $2.50/$15.00 (Source: developers.openai.com), though newer models generally require fewer tokens to reach a comparable answer quality, which lowers effective cost per task even when the per-token price is similar.

What does Llama 3 inference cost compared to proprietary APIs? Llama 3.3 70B costs $1.04 per million tokens (flat input/output rate) on Together AI (Source: www.together.ai, $0.59/$0.79 on Groq (Source: groq.com), and $0.10/$0.32 through OpenRouter (Source: openrouter.ai), each substantially below Claude or GPT-tier pricing, reflecting both the model's smaller effective compute footprint and the competitive dynamics of a market with several providers hosting the identical open-weight model. For historical context, AWS Bedrock's legacy listing for the older Llama 2 Chat 70B model shows $1.95 per million input tokens and $2.56 per million output tokens (Source: aws.amazon.com, nearly double Llama 3.3 70B's current Together AI rate despite Llama 2 being a smaller, less capable, and older model, illustrating how quickly open-weight hosting prices have fallen even generation-over-generation within the same model family.

Is self-hosting an LLM on GPUs cheaper than using an API? It depends entirely on volume and utilization. At low volume, a rented H100 costing $2.89 to $3.99 per hour (Source: www.runpod.io sitting idle costs more than an API subscription would. At high, sustained volume, a named real-world migration documented by LLMDeploy cut a client's monthly cost by 80 percent (Source: llmdeploy.to), and a modeled teaching scenario projects a 97 percent reduction for a similar migration (Hypothetical Example) (Source: academy.talki-app.fr, though both figures should be read as illustrative rather than as independently audited comparisons.

Why do LLM prices keep falling? Multiple independent analyses attribute the decline primarily to software and algorithmic efficiency gains rather than cheaper hardware. The 2026 "Tiered Super-Moore's Law" analysis found GPU hardware contributed only -0.9 percent to observed cost reductions, with total factor productivity residuals, essentially better software, architectures, and competition, accounting for roughly 103.7 percent

(Source: arxiv.org). a16z separately points to quantization, smaller and better-trained models, software optimization, and open-source competition as the main drivers (Source: a16z.com).

How much does it cost to rent a GPU to run an LLM? As of July 2026, on-demand H100 GPUs rent from roughly $1.38 to $12.29 per GPU-hour depending on provider, with neocloud specialists like RunPod, Lambda, and CoreWeave clustering around $2.89 to $6.16 per hour and hyperscalers charging a substantial premium (Source: deploybase.ai).

What is DeepSeek's inference cost compared to GPT-4o or Claude? DeepSeek's official API prices its flagship chat model at $0.14 per million input tokens on a cache miss and just $0.28 per million output tokens (Source: api-docs.deepseek.com), roughly 18 times cheaper on input and 36 times cheaper on output than GPT-4o's $2.50/$10.00 per million (Source: artificialanalysis.ai, and more than 35 times cheaper on input than Claude Opus 4.8's $5.00 per million (Source: www.anthropic.com). Hosted through AWS Bedrock rather than DeepSeek's own API, the same DeepSeek V3.2 model costs more, $0.62 input and $1.85 output per million tokens, reflecting Bedrock's typical premium over a first-party API (Source: aws.amazon.com).

Does prompt caching meaningfully change the cost per million tokens? Yes, substantially for workloads with repeated context. Anthropic's cache-read pricing for Claude Opus 4.8 is $0.50 per million tokens against a $5.00 base input rate, a 90 percent discount for previously processed prompt content (Source: www.anthropic.com), while Google's Gemini context-caching price for Gemini 2.5 Pro is $0.225 to $0.45 per million tokens depending on prompt length, alongside a separate hourly storage fee (Source: ai.google.dev). Artificial Analysis notes that caching mechanics differ enough by provider, Anthropic charges a separate cache-write fee while Google adds an hourly storage charge and OpenAI and DeepSeek generally do not, that buyers comparing sticker prices across vendors need to model their actual cache-hit rate rather than comparing base list prices alone (Source: artificialanalysis.ai).

Conclusion

Cost per million tokens in mid-2026 is best understood not as a single number but as a landscape shaped by three converging forces: an aggressive commodity-tier price war pushing input token costs toward fractions of a cent, a defended frontier-reasoning tier where prices have barely moved despite the broader trend, and a GPU cloud market whose own price declines and throughput gains are steadily lowering the volume threshold at which self-hosting outcompetes a managed API. The specific figures in this report, DeepSeek's $0.0028 cache-hit rate, Anthropic's $2 to $5 Sonnet and Opus tiers, OpenAI's $30 to $180 frontier reasoning pricing, and GPU rentals spanning $1.38 to over $12 per hour for identical H100 hardware, will all be different, and very likely lower, within another twelve months, consistent with decline rates that every research group examined in this report, from a16z to Epoch AI to the peer-reviewed economic literature, measured in multiples per year rather than single-digit percentages.

For a team making a build-versus-buy decision today, the evidence assembled here supports three concrete conclusions. First, the price gap between commodity and frontier models has grown wide enough that routing architectures, sending only genuinely hard tasks to expensive models, now offer a larger cost-optimization lever than choosing any single provider. Second, self-hosting economics improve fastest for organizations processing well over a hundred million tokens per month and willing to accept the operational burden of GPU provisioning, model versioning, and serving-stack optimization; the documented migrations in this report cut costs by 80 to 97 percent, but only after absorbing multi-day engineering efforts and, in some cases, a compliance requirement that made the decision non-negotiable regardless of raw economics. Third, because every reputable price-decline analysis examined here found the cost curve still falling, and in some segments accelerating, any cost model built today should be revisited on a recurring basis rather than treated as a fixed input, since the cheapest viable option for a given workload is, more likely than not, going to change again before the year is out.

Tags: llm cost per million tokens, llm inference pricing, gpu inference cost comparison, llm token cost comparison, self hosted llm vs api cost, gpt-4o pricing, llama 3 inference cost, llm price performance benchmark

---