Hosted AI APIs vs Self-Hosted Models: Cost Comparison

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · hosted ai api · self-hosted llm · llm cost comparison · gpu inference cost · total cost of ownership · ai infrastructure · api pricing · on-premise ai deployment · build vs buy ai · gpu cloud pricing

Executive Summary

Worldwide spending on AI infrastructure reached $89.9 billion in the fourth quarter of 2025 alone, capping a full-year 2025 total of $318 billion and a forecast climb past $1 trillion by 2029 at roughly a 31% five-year compound annual growth rate (idc.com) (idc.com). A large share of that spend traces back to one recurring choice: whether a workload calls a hosted AI API billed per token, or runs on self-hosted GPU capacity billed per GPU-hour. This report, current as of September 2026, builds a transparent, reproducible model for that decision rather than asserting a single universal answer, because the evidence gathered here shows none exists.

Hosted API pricing spans roughly $0.20 to $10 per million tokens (MTok) of input and $1.20 to $50 per MTok of output across the model tiers surveyed from Anthropic, OpenAI, Google, Mistral, and Cohere, with every major vendor offering a roughly 50% batch-processing discount and some offering cache discounts up to 90% (anthropic.com) (openai.com) (mistral.ai). Self-hosted GPU rental rates for the same NVIDIA H100 accelerator range from $2.69 per hour on RunPod's Community Cloud to $12.29 per hour (equivalent, per third-party tracking)

on Microsoft Azure, a four-to-fivefold spread before any question of utilization is considered (runpod.io) (instances.vantage.sh). A published workload-based model finds a 17.5–36.3× cost change within an unchanged H100 configuration when offered request rate changes, while its $0.21–$15.25 per-million-output-token range spans models, precision settings, and GPU allocations (arxiv.org). At one request per second, its two-GPU Mixtral 8x7B FP16 result is $15.25/MTok; at peak load, its one-GPU Llama 3.1 8B results are 6,238 tok/s and $0.311/MTok for FP16, versus 8,155 tok/s and $0.238/MTok for FP8 (arxiv.org).

Enterprise adoption data reflects this ambiguity rather than resolving it. Menlo Ventures' survey of 495 U.S. enterprise AI decision-makers found enterprise open-source model usage fell from 19% to 11% of total share year over year even as overall infrastructure spend roughly doubled, while a16z's 2026 survey of 100 Global 2000 executives found 80% now comfortable hosting models directly with a frontier lab (menlovc.com) (a16z.com). At the same time, Gartner forecasts $80 billion in worldwide sovereign cloud spending in 2026, and independent measurement studies of production GPU clusters, including one covering more than 20,000 GPUs, found utilization commonly sits below 60% and sometimes below 50%, a gap that directly inflates self-hosted cost per token when it occurs (w.media) (arxiv.org).

The report's method section uses effective cost per million output tokens = (GPU hourly price × 1,000,000) / (aggregate output tokens per second × 3,600), so a reader can substitute measured throughput and provider pricing while accounting for model, precision, offered load, and latency requirements (arxiv.org). Staffing, power, and facility overhead (a specialized GPU infrastructure engineer estimated at $275,000 per year fully loaded (introl.com), and data-center power overhead averaging a 1.56 Power Usage Effectiveness ratio industry-wide (datacenter.uptimeinstitute.com)) further shift the calculation for organizations weighing outright hardware ownership against rental or a hosted API. No single break-even point applies across models, providers, and workloads; the evidence instead points toward a blended sourcing strategy shaped by request volume, data-residency requirements, and task-specific quality parity between hosted and open-weight models.

Introduction and Background

Enterprise spending on artificial intelligence infrastructure crossed $318 billion worldwide in 2025 and is forecast to exceed $1 trillion by 2029, a roughly 31% five-year compound annual growth rate (idc.com) (idc.com). A large share of that spend now goes toward a single recurring decision: whether a workload should call a hosted large language model (LLM) application programming interface (API), such as those sold by OpenAI, Anthropic, or Google, or whether it should run on infrastructure the organization rents or owns outright, most commonly on graphics processing units (GPUs) such as NVIDIA's H100, H200, or B200 accelerators.

This report builds a transparent, reproducible cost model for that decision as of September 2026. It does not claim a universal break-even point between hosted and self-hosted inference, because none exists: the answer depends on request volume, achieved GPU utilization, staffing overhead, model choice, and how an organization values latency, data control, and operational risk. Instead, this report documents the inputs, publishes the method, and works through named scenarios so a reader can substitute their own workload figures and reach their own number.

The comparison covers inference (running a trained model to answer requests), not the substantially larger cost of pretraining a foundation model from scratch. It focuses on the two dominant sourcing paths for LLM inference: **hosted API** consumption billed per token, and **self-hosted deployment** on rented or owned GPU capacity running an open-weight model through a serving framework such as vLLM, NVIDIA's TensorRT-LLM, or SGLang. A

companion IntuitionLabs analysis, H100 Rental Prices: A 2026 Cloud Comparison, surveyed rental rates across a wider set of cloud GPU providers; this report does not restate that survey and instead builds on it with a distinct, workload-based cost method, newer pricing observations, and a direct comparison against current hosted API rates. IntuitionLabs states that its Private LLM Inference service runs the application on infrastructure it controls on a customer's behalf.

Hosted AI APIs

Capabilities

Hosted API pricing is published per million tokens (MTok) of text processed, split between input (prompt) and output (completion) tokens, with output tokens consistently priced several times higher than input tokens across every vendor surveyed. As of September 2026, **Anthropic** prices its flagship **Claude Opus 5** at $5 per MTok input and $25 per MTok output, with the mid-tier **Claude Sonnet 5** at $2/$10 per MTok and a cached-input read rate of $0.20/MTok (anthropic.com) (anthropic.com). **OpenAI** prices its top model, **GPT-6 Astra**, at $10/MTok input and $50/MTok output, while the mid-tier **GPT-5.6 Sol** carries promotional pricing of $4/$20 per MTok (down from $5/$30) through at least November 21, 2026, with a cheaper **GPT-5.6 Terra** at $2/$12 and the smallest **GPT-5.6 Luna** at $0.20/$1.20 (openai.com) (openai.com). **Google's Gemini 3.1 Pro Preview** lists at $2.00/MTok input and $12.00/MTok output for prompts under 200,000 tokens, rising to $4.00/$18.00 above that context length (ai.google.dev). **Mistral's** flagship **Medium 3.5** prices at $1.5/$7.5 per MTok, with the smaller **Small 4** at $0.15/$0.6 (mistral.ai), and **Cohere's** most recently published **Command R+** tier (08-2024 revision) lists at $2.50/$10.00 per MTok, itself a cut from an earlier $3.00/$15.00 (cohere.com).

Every major vendor discounts asynchronous, delayed-response workloads. Anthropic, OpenAI, Google, and Azure OpenAI Service each publish a batch-processing discount of 50% off standard token pricing for jobs that do not require an immediate response (anthropic.com) (openai.com) (ai.google.dev) (azure.microsoft.com), and Mistral separately discounts cached (repeated) input tokens by 90% (mistral.ai). **Amazon Bedrock** lists on-demand pricing for the legacy Claude 3.5 Sonnet model at $6.00/MTok input and $30.00/MTok output effective December 1, 2025, with the same 50% batch discount applied, and markets an "Intelligent Prompt Routing" feature that automatically shifts requests within a model family to cut cost by up to 30% without a published accuracy penalty (an AWS-stated claim, not independently benchmarked here) (aws.amazon.com) (aws.amazon.com). **Azure OpenAI Service** offers a parallel path for large, steady workloads: Provisioned Throughput Units (PTUs) reserved at a 15-PTU minimum, priced at $1 per PTU-hour on-demand, $260 per PTU-month, or $2,652 per PTU-year under longer commitments, a structure that converts variable per-token cost into a fixed capacity reservation once volume is predictable (azure.microsoft.com).

Rate limits, not just price, shape effective cost at scale. Anthropic publishes explicit per-model caps across requests-per-minute (RPM), input-tokens-per-minute (ITPM), and output-tokens-per-minute (OTPM); on its "Build" tier, Claude Opus 5 is capped at 1,000 RPM, 2,000,000 ITPM, and 400,000 OTPM (docs.claude.com) (docs.claude.com). OpenAI instead ties rate-limit tier upward mobility to cumulative account spend, automatically raising caps as an account crosses spend thresholds (platform.openai.com). A workload that needs sustained throughput above these published ceilings must negotiate a custom enterprise agreement, a step several vendors gate behind volume-based discounts (Google's Gemini Enterprise tier explicitly offers "volume-based discounts

(based on usage)" on top of standard pricing) (ai.google.dev), or behind an enterprise premium (Mistral prices its Enterprise APIs, with regional data controls and higher rate limits, at a 75% premium over list pricing on select models) (mistral.ai).

Adoption

Enterprise spending on generative AI infrastructure, including hosted model APIs, reached an estimated $18 billion in 2025, roughly half of total enterprise generative AI spend tracked by venture firm Menlo Ventures in a survey of 495 U.S. enterprise AI decision-makers fielded in November 2025 (menlovc.com) (menlovc.com). The same survey found enterprise reliance on open-source models fell year over year, from a 19% share of enterprise LLM usage to 11% (menlovc.com), a shift consistent with continued reliance on hosted proprietary APIs rather than self-hosted open-weight alternatives. Separately, a16z's early-2026 survey of 100 verified vice president and C-level executives at Global 2000 companies found 80% of enterprises now say they are comfortable letting a frontier AI lab host their models directly, rather than routing exclusively through a cloud service provider (CSP) intermediary, alongside 65% who said they still prefer incumbent AI vendors over newer alternatives, citing trust and integration (a16z.com) (a16z.com) (a16z.com).

Strengths and Limitations

The hosted API model's principal strength is speed to deployment: an organization can send its first production request within hours, with no GPU procurement, no serving-framework tuning, and no capacity planning. Cost scales with usage in a way that requires no fixed capital outlay, and batch and cache discounts materially lower cost for workloads that tolerate delay or repetition. The published rate-limit and usage-tier structures, however, mean an organization forecasting sustained high-volume traffic cannot simply assume list pricing will hold. It must budget for either an enterprise agreement, a provisioned-throughput commitment such as Azure's PTUs, or a routing strategy across model tiers, each of which reintroduces some of the planning overhead the hosted model is meant to avoid.

Self-Hosted LLM Infrastructure

Capabilities

Self-hosting substitutes a per-token vendor price for a GPU-hour price, payable either as cloud rental or on-premise ownership. Rental rates vary sharply by provider and GPU generation as of September 2026. On **RunPod**, an H100 SXM instance rents for $2.69/hour (Community Cloud) to $3.29/hour (Secure Cloud), an H200 for $3.59 to $4.59/hour, and a B200 for $5.98 to $6.79/hour (runpod.io). **Lambda Labs** prices an 8-GPU H100 SXM cluster instance at $3.99 per GPU-hour and an 8-GPU B200 SXM6 cluster at $6.69 per GPU-hour (lambda.ai) (lambda.ai). **CoreWeave's** 8-GPU HGX nodes list at $49.24/hour for H100 (about $6.16 per GPU-hour), $50.44/hour for H200 (about $6.31 per GPU-hour), and $68.80/hour for B200 (about $8.60 per GPU-hour) (coreweave.com). Hyperscaler pricing runs higher still: **Google Cloud's** A3 Ultra instance (8x H200) lists at $84.81/hour, about $10.60 per GPU-hour (cloud.google.com), while third-party cloud-pricing tracker Vantage reports **AWS** EC2 P5.48xlarge (8x H100) at $55.04/hour on-demand and **Azure's** ND96isr H100 v5 (8x H100) at $98.32/hour on-demand, both well above the specialized-provider rates (instances.vantage.sh) (instances.vantage.sh). IntuitionLabs' own prior survey of H100 rental listings found a wider spread still, from $1.49/hour on the peer-to-peer marketplace Vast.ai up to

$6.98/hour on Azure (intuitionlabs.ai), and separately found that AWS's own P5 on-demand pricing had fallen to roughly $3.90 per GPU-hour following 2025 price cuts (intuitionlabs.ai), both observations dated to that report rather than re-verified here.

Buying hardware outright shifts the cost from a recurring rate to a capital purchase and a depreciation schedule. Market-tracking estimates put a single new H100 80GB card at roughly $31,000 and a complete 8-GPU HGX H100 server system at $250,000 to $320,000 as of an August 2026 verification pass (cloudzero.com). IRS Publication 946 states that, unless the taxpayer elects out, it must take a 100% special depreciation allowance for certain qualified property acquired and placed in service after January 19, 2025 (irs.gov).

Power draw and facility overhead add a further layer of cost that a rental rate typically already prices in, but that an on-premise buyer must model separately. NVIDIA's own specifications list the H100 SXM module at up to 700 watts (W) of configurable thermal design power (TDP), against 350 to 400W for the PCIe card variant (nvidia.com), and the H200 SXM at the same 700W ceiling against up to 600W for its PCIe/NVL form factor (nvidia.com). The Uptime Institute's 2024 Global Data Center Survey, based on 879 respondents, found the industry-average Power Usage Effectiveness (PUE), the ratio of total facility power draw to power actually delivered to computing equipment, at 1.56, and noted that average PUE has remained essentially flat for five consecutive years across the industry's large base of older facilities (datacenter.uptimeinstitute.com) (uptimeinstitute.com). A vendor cost model published by GPU-hosting firm Introl illustrates the combined effect: at an assumed PUE of 1.5, a 100-GPU cluster drawing 400 kilowatts (kW) of GPU power draws 600kW at the meter, and the same model estimates U.S. colocation space at, separately, roughly $196 to $250 per kW per month based on a CBRE market report covering the second half of 2025 (introl.com) (inflect.com).

Running the model efficiently once hardware is secured depends on the serving software layer. Open-source frameworks including **vLLM**, **NVIDIA's TensorRT-LLM**, **Hugging Face's Text Generation Inference (TGI)**, and **SGLang** manage GPU memory and request batching to raise achievable throughput per GPU; vLLM's original PagedAttention paper reported a 2 to 4 times throughput improvement over then-prior serving systems at equivalent latency, and up to 22 times higher sustainable request rates than the earlier FasterTransformer system (arxiv.org) (arxiv.org).

Adoption

Self-hosting and on-premise deployment remain concentrated among larger enterprises with heavier compliance exposure. A16z's 2025 CIO-focused survey found open-source model adoption was occurring disproportionately "at the larger end of enterprises where on-prem is still a major consideration" (a16z.com), and Gartner has forecast that worldwide sovereign cloud infrastructure-as-a-service (IaaS) spending, capacity kept within a jurisdiction's own borders or control for data-sovereignty reasons, will reach $80 billion in 2026, up 35.6% year over year, shifting an estimated 20% of workloads from global to local providers (w.media) (w.media). Cost, not only compliance, is pulling some workloads back on-premise generally: Flexera's 2025 State of the Cloud survey found respondents had moved more than one-fifth of their cloud workloads back on-premise, a cloud-repatriation trend not specific to AI workloads but consistent with the same cost logic (ciodive.com). Separately, Deloitte's State of Ethics and Trust in Technology survey of over 1,800 professionals found data privacy ranked as the top technology concern for 40% of respondents, a driver frequently cited for keeping model inference inside an organization's own infrastructure boundary (prnewswire.com).

Strengths and Limitations

Self-hosting's principal strength is control: an organization sets its own data-residency boundary, chooses its own model weights (including fine-tuned or domain-adapted variants), and is not subject to a vendor's rate-limit ceiling. Its principal cost risk is underutilization. A widely cited measurement study of ByteDance's production deep-learning cluster, spanning more than 20,000 GPUs, found GPU compute and streaming-multiprocessor utilization stayed below 60% on more than 99% of GPUs measured (arxiv.org), and a separate 2024 academic survey of cluster-scheduling research reported that recorded production GPU utilization "typically ranges from 25% to below 50%" (arxiv.org). Because a rented or owned GPU is billed (or depreciated) whether or not it is processing a request, low utilization directly inflates the effective cost per token, the mechanism explored quantitatively in the Data Analysis section below. Self-hosting also carries a fixed staffing cost that scales weakly with usage: a GPU-infrastructure-focused cost model published by hosting firm Introl estimates a specialized GPU infrastructure engineer at $275,000 per year fully loaded, and separately states that hardware acquisition represents only about 35% of a cluster's five-year total cost of ownership, implying the remainder is power, networking, facilities, and staffing (introl.com) (introl.com). The U.S. Bureau of Labor Statistics' (BLS) closest tracked occupational proxies, Software Developers (May 2025 data) and Network and Computer Systems Administrators (May 2024 data, the most recent published at the time of access), reported median annual pay of $135,980 and $96,800 respectively (web.archive.org) (web.archive.org), and aggregator Levels.fyi separately reports an average total compensation of $280,000 for a Machine Learning Engineer title as of September 2026, though this is a self-reported, non-government data source and should be read as directional rather than authoritative (levels.fyi).

Feature Comparison

Table 1 below summarizes the qualitative trade-offs between hosted API consumption and self-hosted GPU deployment across the dimensions most relevant to an infrastructure sourcing decision, drawing on the vendor documentation and market data cited above.

Dimension	Hosted AI API	Self-Hosted (Rented or Owned GPU)
Pricing unit	Per-million-token, split input/output, e.g. $2 to $10 per MTok for Anthropic's Claude Sonnet 5 (anthropic.com)	Per-GPU-hour rental (e.g. $2.69 to $8.60/hr for H100 to B200 across providers surveyed) (runpod.io) (coreweave.com) or amortized purchase cost
Time to first production request	Hours; no capacity planning required	Days to weeks; requires provisioning, serving-framework setup, and model selection
Cost driver at low volume	Scales down to near-zero at idle	Fixed regardless of request volume; GPU-hour or depreciation cost accrues whether utilized or not
Cost driver at high, steady volume	List price plus rate-limit or enterprise-tier negotiation (platform.openai.com) (ai.google.dev)	Amortized GPU-hour cost falls as utilization rises toward the framework's achievable throughput ceiling (arxiv.org)
Data residency / control	Data processed on vendor infrastructure, subject to vendor terms	Full control of data location and model weights; supports sovereign/on-premise requirements (w.media)

Dimension	Hosted AI API	Self-Hosted (Rented or Owned GPU)
Staffing requirement	Minimal; vendor manages serving infrastructure	Requires GPU infrastructure engineering capability, estimated at $275,000/year fully loaded per specialist role (introl.com)
Model choice	Limited to vendor's published model family	Open to any open-weight model (Llama, Mixtral, DeepSeek, GLM, etc.) and fine-tuned variants
Discounting mechanisms	Batch (50% typical), prompt caching (up to 90% on Mistral), provisioned throughput commitments (mistral.ai) (azure.microsoft.com)	Reserved/committed-use cloud discounts, spot pricing (CoreWeave spot rates run roughly 40 to 60% below on-demand) (coreweave.com)

The table's central asymmetry is utilization sensitivity. A hosted API's per-token price is fixed regardless of how continuously an organization sends requests, so its total cost scales almost linearly with usage. A self-hosted deployment's per-token price is not fixed at all: it is a function of how close actual request volume comes to the GPU's achievable throughput ceiling, which is why the same hardware can be dramatically cheaper or more expensive than a hosted API depending on workload shape, a relationship quantified in the Data Analysis section below.

Performance and Benchmarks

Independent throughput benchmarks show a wide, fast-moving spread across serving frameworks and GPU generations. NVIDIA's own TensorRT-LLM documentation reports an H100 running FP8 (8-bit floating point) precision sustaining over 10,000 output tokens per second at peak throughput while holding a 100-millisecond first-token latency budget, and states that H100 with TensorRT-LLM delivers up to 4.6 times the maximum throughput of the prior-generation A100 on comparable workloads (nvidia.github.io) (nvidia.github.io). NVIDIA's published performance tables show a single H100 running an 8-billion-parameter Llama 3.1 model at FP8 precision exceeding 26,400 tokens per second at short input/output lengths (nvidia.github.io), and a separate NVIDIA benchmark for the larger Mixtral 8x7B model on two H100 GPUs reports 38.4 requests per second sustained throughput (developer.nvidia.com). Independent comparative testing published on arXiv in November 2025 found vLLM achieving up to 24 times higher throughput than Hugging Face's TGI framework under high-concurrency conditions, with a concrete measurement on Llama-2-7B of 15,243 tokens per second for vLLM against 4,156 tokens per second for TGI at 100 concurrent requests (arxiv.org) (arxiv.org). The SGLang framework's own developers at LMSYS report SGLang achieving up to 3.1 times higher throughput than vLLM specifically on Llama-70B serving, with both SGLang and TensorRT-LLM reaching up to 5,000 tokens per second on short-input workloads in their published comparison (lmsys.org) (lmsys.org). These figures come from each framework's own benchmark methodology and are not independently re-verified against a single common test harness here; a reader building a serving decision on throughput alone should treat cross-framework comparisons as directional rather than definitive, since benchmark configuration (batch size, sequence length, precision) materially changes the result.

Quality, not only throughput, factors into any hosted-versus-self-hosted decision, because a cheaper self-hosted deployment is only a genuine substitute if the open-weight model it runs performs comparably on the target task. The LMArena Text leaderboard (formerly Chatbot Arena), an independently operated, crowd-voted comparative ranking, showed as of September 2, 2026 that the open-weight, MIT-licensed DeepSeek-V4-Pro-high model scored

1,460 (plus or minus 8), close behind the top-ranked proprietary Claude Fable 5 model's score of 1,507 (plus or minus 5) (lmarena.ai). The same leaderboard placed the open-weight, MIT-licensed GLM-5.3-max model (from Z.ai), priced at $1.40/$4.40 per MTok when accessed via API, at rank 20 overall with a score of 1,482 (plus or minus 7), ahead of several proprietary models ranked lower on the same board (lmarena.ai). These rankings reflect a specific, crowd-sourced comparative methodology and should be read as one data point on model quality, not a substitute for task-specific evaluation on an organization's own workload.

An arXiv-published cost-efficiency study found that replacing OpenAI's API with self-hosted open-source small language models (SLMs) yielded a cost reduction of 5 to 29 times, with the exact multiple depending heavily on which open-source model was substituted and at what accuracy parity (arxiv.org). That wide range is itself evidence against a single break-even claim: the achievable saving depends on model selection, task difficulty, and (as the next section quantifies directly) utilization.

Data Analysis and Evidence

Method

This section applies a workload-based model from a 2026 arXiv paper. For output tokens, effective cost per million tokens is (GPU hourly price × 1,000,000) / (aggregate output tokens per second × 3,600). The paper parameterizes cost by hardware, model architecture, quantization precision, offered request rate, and latency/SLO conditions; a reproducible estimate therefore requires each of those configuration-specific inputs, rather than treating one H100 result as universal (arxiv.org).

The cited paper's full $0.21–$15.25/MTok range covers six H100 configurations and their load ranges, not utilization alone. Its within-configuration near-idle-to-saturation penalty is 17.5× for Mixtral 8x7B FP16 through 36.3× for Qwen3-30B-A3B FP8. The $15.25 point is Mixtral 8x7B FP16 on two H100s at one request per second; for one-GPU Llama 3.1 8B at peak load, FP16 reaches 6,238 tok/s at $0.311/MTok and FP8 reaches 8,155 tok/s at $0.238/MTok (arxiv.org).

Worked Scenarios

Table 2 summarizes configuration-specific findings from the cited benchmark; it is not a same-hardware, utilization-only comparison.

Configuration and load	GPU allocation and precision	Reported result
Mixtral 8x7B at 1 rps	Two H100s, FP16	$15.25/MTok
Mixtral 8x7B at saturation	Two H100s, FP16	$0.871/MTok; 17.5× lower than at 1 rps
Llama 3.1 8B at peak load	One H100, FP16	6,238 tok/s; $0.311/MTok
Llama 3.1 8B at peak load	One H100, FP8	8,155 tok/s; $0.238/MTok
Qwen3-30B-A3B at peak load	One H100, FP8	$0.209/MTok

Within an unchanged configuration, offered request rate drives the cited 17.5–36.3× near-idle-to-saturation effect. The paper's $0.21–$15.25/MTok range instead crosses configurations with different models, precision settings, and GPU allocations (arxiv.org).

Table 3 places the underlying GPU rental rates behind this model side by side, since the achievable floor of the "effective cost per token" calculation is set by the cheapest available compute for a given GPU generation.

Provider	H100 (per GPU-hour)	H200 (per GPU-hour)	B200 (per GPU-hour)
RunPod (Community / Secure) (runpod.io)	$2.69 / $3.29	$3.59 / $4.59	$5.98 / $6.79
Lambda Labs (8-GPU cluster) (lambda.ai)	$3.99	not published at time of access	$6.69
CoreWeave (8-GPU HGX, on-demand) (coreweave.com)	~$6.16	~$6.31	~$8.60
AWS (P5.48xlarge, on-demand, per third-party tracker)	$6.88 ($55.04 / 8 GPUs) (instances.vantage.sh)	not applicable (different instance family)	not applicable
Google Cloud (A3 Ultra, per-GPU)	not applicable (different instance family)	~$10.60 (cloud.google.com)	not published at time of access
Azure (ND96isr H100 v5, per third-party tracker)	$12.29 ($98.32 / 8 GPUs) (instances.vantage.sh)	not applicable	not published at time of access

The spread across Table 3, roughly $2.69 to $12.29 per H100 GPU-hour depending on provider, means the "GPU rental basis" input to the workload cost model in Table 2 is itself a four-to-fivefold variable before utilization is even considered. A reader reproducing this model should treat both variables, provider rate and achieved utilization, as inputs to substitute with their own measured figures rather than accept either the low or high end of the ranges shown here as a default assumption. Enterprise infrastructure spending data corroborates that this is a live, growing decision rather than a settled one: IDC's Q4 2025 tracker recorded worldwide AI infrastructure spending of $89.9 billion in that quarter alone, capping a full-year 2025 total of $318 billion and more than double 2024 spending (idc.com), spending IDC forecasts will grow at roughly a 31% five-year compound annual growth rate through 2029 (idc.com).

Implications and Future Directions

Several trends visible in the data above will shift the calculus in both directions over the next planning cycle. On the hosted side, per-token prices have already fallen within the period surveyed: the promotional mid-tier pricing described in the Hosted AI APIs section above ran roughly 20 to 33% below the prior list price at the time of this report, and Cohere's Command R+ price fell from $3.00/$15.00 to $2.50/$10.00 per MTok between two published revisions (cohere.com), a pattern consistent with broader competitive pressure on API pricing. Batch, caching, and provisioned-throughput mechanisms described above are also expanding, which lowers the effective hosted price for organizations willing to restructure workloads (tolerating delay, or committing to steady volume) to capture them (azure.microsoft.com).

On the self-hosted side, serving-framework efficiency continues to improve (vLLM's original 2 to 4 times throughput gain over prior systems, and subsequent frameworks each claiming further multiples over vLLM in specific configurations) (arxiv.org) (lmsys.org), which pushes the achievable-throughput ceiling upward and, per the workload model in the Data Analysis section, pushes the effective cost per token downward for any organization able to sustain higher utilization. Open-weight model quality is also narrowing the gap against proprietary hosted models on independent leaderboards, which matters directly to the "comparable task quality" condition underlying any cost comparison: a cheaper self-hosted option is not a genuine substitute if it cannot perform the target task (lmarena.ai) (lmarena.ai).

Regulatory and data-sovereignty pressure is pulling in the same direction as cost for a specific subset of workloads. Gartner's forecast of $80 billion in sovereign cloud IaaS spending in 2026, and its estimate that 20% of workloads will shift from global to local providers, describes infrastructure generally rather than AI inference specifically, but the same jurisdictional logic applies with particular force to regulated industries handling patient, clinical, or other sensitive data, where data residency requirements can make self-hosting (or a sovereign-cloud hosted variant of it) a compliance requirement rather than only a cost optimization (w.media) (w.media). For life-sciences and other regulated organizations navigating this specific intersection of cost, utilization, and compliance, a scenario-specific model can translate a workload's actual request pattern and data-governance requirements without imposing a single default answer.

Enterprises appear to be converging on a blended answer rather than an exclusive one. Menlo Ventures' finding that enterprise open-source model usage fell from 19% to 11% of total usage share year over year, even as overall AI infrastructure spend roughly doubled, suggests many organizations are consolidating routine, variable-volume workloads onto hosted APIs while reserving self-hosted or on-premise capacity for the specific workloads where volume, latency, data-residency, or customization requirements justify the fixed cost and staffing commitment (menlovc.com) (idc.com).

Frequently Asked Questions (FAQs)

Is there a single request-volume threshold at which self-hosting becomes cheaper than a hosted API? No single threshold applies across models and providers. The cited workload-based cost model finds a 17.5–36.3× within-configuration cost change as offered request rate changes; its $0.21–$15.25/MTok range spans six configurations with different models, precisions, and GPU allocations (arxiv.org), and the GPU rental rate itself varies roughly four to fivefold across providers for the same H100 GPU generation, per Table 3 above (instances.vantage.sh). An organization must model its own achievable throughput and its own available rental rate to find its own break-even point.

When does self-hosting make sense regardless of raw cost? Survey evidence points to two recurring, non-cost drivers: data-residency or compliance requirements, cited by 40% of respondents in Deloitte's technology-trust survey as their top technology concern (prnewswire.com), and latency-sensitive or high-volume steady workloads, cited by Menlo Ventures as a driver of continued on-device and edge deployment (menlovc.com). A16z's CIO survey similarly found open-source and on-premise adoption concentrated among larger enterprises where these considerations are already "a major consideration" (a16z.com).

How does GPU inference cost compare to API pricing per token? Table 3 in the Data Analysis section lists per-GPU-hour rental rates from $2.69 (RunPod H100, Community Cloud) to $12.29 (Azure ND96isr H100 v5, per third-party tracker) (instances.vantage.sh), while hosted API prices in this report range from $0.20/MTok input (OpenAI GPT-5.6 Luna) to $10/MTok input on the input side, reaching as high as $50/MTok on the output side (openai.com). Direct comparison requires converting the GPU-hour rate into a per-token figure using the workload model above, since the GPU-hour rate alone does not indicate cost per token without a throughput assumption.

What does a workload-based cost model need as inputs to be reproducible? At minimum: the GPU rental rate (or amortized purchase and depreciation cost) per hour, the achieved throughput in tokens per second at the organization's actual, measured request concurrency, the serving framework in use (since framework choice can shift achievable throughput by a reported 2 to 24 times depending on comparison) (arxiv.org) (arxiv.org), and the model's precision (FP8 deployments in the cited studies consistently reached higher throughput, and therefore lower effective cost per token, than FP16 deployments of comparable models) (arxiv.org) (arxiv.org).

Is build-versus-buy for AI infrastructure a one-time decision? Available data suggests it is not. Worldwide AI infrastructure spending nearly doubled from 2024 to 2025 and is forecast to keep growing at roughly 31% annually through 2029 (idc.com) (idc.com), while enterprise open-source usage share fell over the same period even as overall infrastructure spend grew (menlovc.com), indicating organizations are actively re-weighing the hosted-versus-self-hosted mix as pricing, throughput, and model quality all continue to shift.

Conclusion

Hosted AI APIs and self-hosted GPU infrastructure represent two different cost structures rather than two points on a single price scale. A hosted API's cost scales close to linearly with token volume and requires no infrastructure staffing, while a self-hosted deployment's cost is fixed per GPU-hour and its effective cost per token falls as achieved utilization rises toward the serving framework's throughput ceiling. The published H100 study finds a 17.5–36.3× within-configuration cost change as offered request rate varies; its broader $0.21–$15.25/MTok range covers six configurations with different models, precision settings, and GPU allocations (arxiv.org).

No universal break-even point follows from this, and none should be assumed. An organization evaluating the decision needs its own measured request pattern, its own achievable throughput under its chosen serving framework and precision, its own available GPU rental rate (which itself varies by a factor of four to five across the providers surveyed here), and an honest accounting of staffing, power, and facility overhead if it is weighing outright ownership rather than rental. Regulatory, data-residency, and task-quality considerations, not cost alone, will continue to push some workloads toward self-hosting and others toward hosted APIs regardless of which is nominally cheaper on paper, and enterprise survey data suggests most organizations are settling on a blended mix rather than an exclusive choice between the two paths.
