DeepSeek R1 Self-Hosting Cost per Million Tokens (2026 Guide)

Published July 21, 2026 29 min read

Executive Summary

Self-hosting **DeepSeek R1** is not a single price point, it is a hardware decision that happens to produce a cost per million tokens. The full 671 billion parameter model, of which only 37 billion parameters activate per token in its Mixture-of-Experts (MoE) design (Source: arxiv.org), needs roughly 700GB of memory in native FP8 precision or about 1.4TB if converted to BF16 (Source: theriseunion.com) (Source: theriseunion.com). The vendor's own deployment guidance calls for **8 NVIDIA H200 or H20 GPUs (141GB each)** or **16 H100 or H800 GPUs (80GB each)** for FP8 inference, and 16 A100 or A800 GPUs if BF16 conversion is required (Source: theriseunion.com) (Source: theriseunion.com).

Benchmarked, independently measured cost per million tokens for self-hosted inference varies enormously with batching strategy. SemiAnalysis's InferenceX benchmarking tool puts DeepSeek R1 inference on **NVIDIA H100** at **$1.365 per million tokens** in a low-latency, low-batch configuration, rising to **$4.864** at moderate concurrency and **$15.692** at maximum interactivity per user, while the newer **NVIDIA B200** cuts those figures to **$0.113, $0.544, and $1.088** respectively at the same three operating points (Source: inferencex.semianalysis.com). Renting the hardware itself runs from about **$2.89 per hour for an H100 PCIe on RunPod** (Source: www.runpod.io) to **$4.76 per GPU per hour on CoreWeave's HGX H100 nodes** (Source: www.coreweave.com), up to **$55.04 per hour for an 8x H100 AWS p5.48xlarge instance** on-demand (Source: instances.vantage.sh), meaning a dedicated 16-GPU self-hosted cluster for the full model can run into the tens of thousands of dollars per month before any inference has occurred.

Against that backdrop, buying tokens from an **application programming interface (API)** looks inexpensive. DeepSeek's own API originally priced R1 (accessed via the `deepseek-reasoner` model name) at **$0.55 per million input tokens** (cache miss) and **$2.19 per million output tokens**, with cache hits falling to **$0.14 per million tokens** (Source: api-docs.deepseek.com) (Source: api-docs.deepseek.com). Critically, as of the July 22, 2026 publication date of this report, DeepSeek has announced that the `deepseek-chat` and `deepseek-reasoner` model name aliases will be **retired on July 24, 2026**, rerouted to the newer `deepseek-v4-flash` thinking mode (Source: api-docs.deepseek.com), a two-day window from this article's publication that makes the underlying question, whether to keep renting an API alias or to self-host the original open-weight model permanently, more urgent than usual. Third-party hosts of the original R1 weights show wide dispersion: prices vary up to **6.1x across providers**, with DeepInfra the most affordable at **$0.56 per million tokens** and Together AI the most expensive at **$3.40 per million tokens** (Source: artificialanalysis.ai).

The verdict this report reaches is scale-dependent. A documented case study from LiftOff LLC found that self-hosting smaller distilled R1 variants on AWS cost **$414 per month per GPU instance**, dramatically more expensive than **$20 per user per month** for ChatGPT Plus at their under-100-user scale (Source: www.zenml.io). Self-hosting the full 671B model only becomes cost-competitive at high, sustained utilization, precisely the regime where SemiAnalysis benchmarks show B200-class hardware delivering sub-$0.15-per-million-token economics (Source: inferencex.semianalysis.com). This report walks through the API baseline, the GPU hardware math, provider-by-provider pricing, and named real-world deployments to answer, precisely, what it costs to run DeepSeek R1 yourself.

Introduction and Background

DeepSeek R1 is an open-weight large language model released by the Chinese AI startup DeepSeek in January 2025, built on the DeepSeek-V3 base model and trained with large-scale reinforcement learning to produce step-by-step reasoning chains before answering (Source: huggingface.co). Both the model weights and the accompanying code repository are released under the permissive **MIT License**, which explicitly allows commercial use, modification, and distillation into other models (Source: huggingface.co), a commitment DeepSeek reiterated at launch by describing the release as "open for the community to leverage model weights & outputs" (Source: api-docs.deepseek.com). That licensing is precisely why "self-hosting cost per million tokens" is a question worth asking at all: unlike a closed model such as GPT-4 or Claude, there is no contractual requirement to buy tokens from the original vendor. Anyone with sufficient hardware and the technical capacity to run inference software such as vLLM or SGLang can download the weights and operate the model indefinitely, even after DeepSeek itself changes its own hosted product line.

That last point has immediate relevance. As of this report's July 22, 2026 publication date, DeepSeek's official pricing page confirms that the `deepseek-reasoner` model alias, the API name that has always pointed to R1 reasoning behavior, will be deprecated on **July 24, 2026** and rerouted to the "thinking mode" of the newer `deepseek-v4-flash` model (Source: api-docs.deepseek.com). That does not delete the original R1 weights from Hugging Face, where they remain permanently downloadable, but it does mean that anyone relying on DeepSeek's hosted API to get R1-specific behavior will soon be served a different underlying model whether they intend that or not. For teams that specifically need R1's original reasoning traces, whether for reproducibility, benchmarking, or regulatory documentation, self-hosting the archived weights is the only way to guarantee the exact same model continues to run after the API cutover.

The economics of that choice hinge on two separate cost curves that this report treats as distinct: the price of **buying tokens** from DeepSeek's own API or from a third-party host, and the price of **renting or owning the GPU capacity** to run the 671B parameter model (or one of its smaller distilled variants) on one's own infrastructure.DeepSeek's own technical report states the full model needs **2,048 NVIDIA H800 GPUs** for training over roughly two months (Source: dev.to), a figure unrelated to inference but illustrative of the model's scale. Inference is materially cheaper than training but still demands a dedicated multi-GPU cluster for the full-size model, which is the central hardware constraint this report quantifies section by section, covering API pricing, GPU requirements, cost-per-million-token math drawn from independently benchmarked data, and named deployments that have actually attempted the self-hosting calculus in production.

DeepSeek R1 API Pricing: The Baseline Before You Self-Host

Before evaluating whether self-hosting makes sense, it is necessary to establish what DeepSeek itself, and the ecosystem of third-party hosts, charge for R1 tokens through a standard API. At its January 2025 launch, DeepSeek priced R1 access (via the `deepseek-reasoner` model parameter) at **$0.14 per million input tokens on a cache hit**, **$0.55 per million input tokens on a cache miss**, and **$2.19 per million output tokens** (Source: api-docs.deepseek.com) (Source: api-docs.deepseek.com). For context on how aggressive that pricing was, OpenAI's directly comparable reasoning model at the time, **o1**, was officially priced at **$15.00 per million input tokens, $7.50 per million cached input tokens, and $60.00 per million output tokens** (Source: developers.openai.com) (Source: developers.openai.com), meaning R1's output pricing undercut o1's by more than **27x** at list price, a gap independent analysis has also confirmed (Source: fireworks.ai).

That original pricing, however, is now a historical artifact of the API rather than a permanent commitment. DeepSeek's current pricing documentation states plainly that the `deepseek-chat` and `deepseek-reasoner` model names will be deprecated on **July 24, 2026 at 15:59 UTC**, with both names remapped to non-thinking and thinking modes of `deepseek-v4-flash` respectively for backward compatibility (Source: api-docs.deepseek.com). Current-generation `deepseek-v4-flash` pricing sits at **$0.14 per million input tokens on a cache miss** and **$0.28 per million output tokens**, with cache hits as low as **$0.0028 per million tokens** (Source: api-docs.deepseek.com), while the larger `deepseek-v4-pro` model runs **$0.435 input (cache miss)** and **$0.87 output** per million tokens (Source: api-docs.deepseek.com). Both current models expose a **1 million token context window** with a maximum output of 384K tokens (Source: api-docs.deepseek.com).

Because the open-weight R1 model remains downloadable regardless of what DeepSeek's own API serves, a secondary market of third-party hosts continues to offer the original architecture at independently set prices, and this is where the "R1 vs GPT-4 cost comparison" and "DeepSeek R1 open source pricing" questions actually get resolved in the market. Artificial Analysis, which benchmarks API providers directly, found **6 providers** actively

serving "DeepSeek R1 0528," the most recent open R1 checkpoint, with blended pricing (a 7:2:1 cache-input-output ratio) ranging from **DeepInfra at $0.56 per million tokens** to **Together AI at $3.40 per million tokens**, a **6.1x spread** (Source: artificialanalysis.ai). DeepInfra's specific input and output rates are **$0.50 per million input tokens** and **$2.15 per million output tokens**, the lowest of the tracked providers (Source: artificialanalysis.ai) (Source: artificialanalysis.ai). Notably, Together AI's current public pricing page no longer lists a standalone DeepSeek R1 entry at all, offering only the newer **DeepSeek V4 Pro** at **$1.74 per million input tokens and $3.48 per million output tokens** (Source: www.together.ai), a sign of how quickly the hosted-model landscape consolidates around newer releases even as the original R1 weights remain independently available. Microsoft Azure's Foundry Models catalog lists **DeepSeek R1 Global at $1.35 per million input tokens and $5.40 per million output tokens** on a pay-as-you-go basis (Source: azure.microsoft.com), and aggregator OpenRouter reports a **weighted average input price of $0.794 and output price of $2.67 per million tokens** across its connected providers over the past 30 days (Source: openrouter.ai) (Source: openrouter.ai). This dispersion, driven by different hardware, quantization, and margin choices among hosts, is the direct market signal that self-hosting is attempting to arbitrage.

GPU Hardware Requirements for Self-Hosting DeepSeek R1

Self-hosting DeepSeek R1 begins with a memory problem before it becomes a cost problem. The full model has **671 billion total parameters, of which 37 billion activate per token** thanks to its Mixture-of-Experts architecture (Source: arxiv.org), but inference still requires the entire parameter set resident in accelerator memory because any token can route to any expert. DeepSeek trained the model natively in **FP8 (8-bit floating point)** precision, so on GPUs with hardware FP8 support, memory usage approximates the raw parameter count: roughly **700GB** for 671 billion parameters (Source: theriseunion.com). On GPUs that lack FP8 tensor cores, such as the widely deployed **NVIDIA A100**, the model must be converted to **BF16 (16-bit brain floating point)**, roughly doubling memory needs to **approximately 1.4TB** (Source: theriseunion.com).

That memory footprint dictates concrete GPU counts. Vendor deployment guidance recommends **8x NVIDIA H200 (141GB) or H20 (141GB)** GPUs, or **16x H100 (80GB) or H800 (80GB)** GPUs, for native FP8 inference of the full model, versus **16x A100 (80GB) or A800 (80GB)** GPUs where BF16 conversion is unavoidable (Source: theriseunion.com) (Source: theriseunion.com). Independent community estimates converge on similar figures: one GPU-requirements guide puts the full model's VRAM need at **approximately 1,543GB** and recommends **16x A100 80GB** GPUs (Source: apxml.com), while another places it at **~1,342GB** with the same 16x A100 recommendation (Source: dev.to). Enterprises standardized on Huawei's Ascend accelerators instead of NVIDIA hardware face a comparable calculation: DeepSeek's full model needs just **2 Atlas 800I A2 servers (8x64GB each)** under W8A8 quantization, or **4 such servers** for unquantized BF16 operation (Source: www.theriseunion.com), an option DeepSeek and Huawei have jointly documented (Source: www.theriseunion.com).

Rental market data corroborates this hardware picture from the demand side. Vast.ai's live marketplace, which aggregates GPU capacity across more than 40 data centers (Source: vast.ai), lists **NVIDIA H100 SXM** units at a median price of **$1.92 per hour** (Source: vast.ai), while **A100 SXM4** units, the workhorse for BF16-only inference of the full R1 model, list at a median of **$0.77 per hour** (Source: vast.ai). CoreWeave's dedicated A100 80GB instances price similarly, at **$2.21 per hour** for either PCIe or NVLINK-connected cards (Source: www.coreweave.com), underscoring that even the cheapest viable full-model configuration, 16 A100 GPUs, carries an hourly hardware bill well into double digits before any inference software or engineering overhead is added.

Table 2 below consolidates verified GPU rental pricing against the vendor-recommended GPU counts for running the full 671B parameter model, translating the hardware requirements above into concrete hourly costs.

GPU	VRAM PER CARD	GPUS NEEDED FOR FULL 671B MODEL	TYPICAL RENTAL PRICE PER GPU PER HOUR
NVIDIA H200	141GB (Source: www.nvidia.com)	8x, FP8 native (Source: theriseunion.com)	$4.39 (RunPod) (Source: www.runpod.io)
NVIDIA H100	80GB (Source: instances.vantage.sh)	16x, FP8 native (Source: theriseunion.com)	$2.89 (RunPod) to $6.88 (AWS on-demand) (Source: www.runpod.io) (Source: instances.vantage.sh)
NVIDIA A100	80GB	16x, BF16 conversion required (Source: theriseunion.com)	$0.77 (Vast.ai median) to $2.21 (CoreWeave dedicated) (Source: vast.ai) (Source: www.coreweave.com)
NVIDIA B200	180GB (Source: lambda.ai)	Not yet specified in vendor deployment guides; delivers lowest benchmarked cost per token (Source: inferencex.semianalysis.com)	$6.69 to $9.86 (Lambda) (Source: lambda.ai)

As *Table 2* shows, the theoretical minimum-hardware entry point for self-hosting the full model starts near $2.89 per GPU hour on commodity H100 clusters, but scales past $4 per GPU hour once enterprise-grade managed platforms or newer accelerators like H200 and B200 are involved. Multiplying these per-GPU rates by the 8 to 16 card counts vendor guidance requires is what produces the tens-of-thousands-of-dollars-per-month figures discussed elsewhere in this report, and it is why the SemiAnalysis-benchmarked cost-per-token figures cited throughout treat GPU generation as the single biggest lever an operator controls.

Quantization meaningfully changes this math. One hardware guide finds that reducing the DeepSeek-R1-Distill-Qwen-32B checkpoint to a full-precision deployment still needs roughly **82GB of VRAM**, achievable on a multi-GPU consumer setup such as **four NVIDIA RTX 4090 cards** rather than a single data-center accelerator (Source: apxml.com), while reducing the full 671B model itself to **4-bit precision** cuts VRAM needs to roughly **436GB**, bringing the requirement down to **6x A100 80GB GPUs** rather than sixteen (Source: apxml.com), though at a documented quality cost that most production teams accept only reluctantly. For teams unwilling to run the full 671B model at all, DeepSeek also released six openly licensed **distilled models**, ranging from **1.5B to 70B parameters**, built by fine-tuning existing Qwen2.5 and Llama3 checkpoints on reasoning traces generated by full R1 (Source: huggingface.co), and DeepSeek's own benchmarks show **DeepSeek-R1-Distill-Qwen-32B outperforming OpenAI's o1-mini across math, code, and reasoning benchmarks** (Source: huggingface.co), a meaningful data point for teams weighing distillation against the full model's hardware burden.

Calculating Self-Hosted Cost per Million Tokens

The central question this report answers, "what does self-hosting cost per million tokens," resolves into a benchmarking problem rather than a simple arithmetic one, because inference cost per token depends heavily on batching and target latency (how many concurrent users a single GPU serves before responses slow down). SemiAnalysis's InferenceX benchmarking platform provides the most direct, independently measured answer available. Interpolated from real benchmark data across three operating points ranging from low to high interactivity per user, **NVIDIA H100** delivers DeepSeek R1 inference at **$1.365, $4.864, and $15.692 per million tokens** as interactivity demands rise, while **NVIDIA B200** achieves **$0.113, $0.544, and $1.088 per million tokens** at the same three points (Source: inferencex.semianalysis.com). The underlying throughput driving those costs ranges from **266.3 tokens per second per GPU on H100** down to **23.3 tokens per second per GPU** at the highest-interactivity setting, versus **4,791.6 down to 497.5 tokens per second per GPU on B200** (Source: inferencex.semianalysis.com). In practical terms, an operator willing to accept slower per-user response times in exchange for higher batch throughput can push the effective cost per million tokens on H100 down close to API-comparable levels, while an operator demanding instant, low-latency responses for every concurrent user pays roughly **11x more per token** on the same hardware.

That benchmarked cost per token has to be paired with the hourly price of the GPUs themselves. Verified on-demand hourly rental rates as of July 2026 span from **RunPod's H100 PCIe at $2.89 per hour** (Source: www.runpod.io) and **H100 SXM at $2.99 per hour** (Source: www.runpod.io), through **Lambda's single H100 SXM at $3.99 per hour** (Source: lambda.ai), to **CoreWeave's H100 PCIe at $4.25 per hour** (Source: www.coreweave.com) and its 8-GPU **HGX H100 node at $4.76 per GPU per hour** (Source: www.coreweave.com). At the high end, **AWS's on-demand p5.48xlarge, an 8x H100 instance, starts at $55.04 per hour**, or roughly **$6.88 per GPU per hour** (Source: instances.vantage.sh), with

one-year reserved pricing dropping to **$23.777 per hour** for the full instance (Source: instances.vantage.sh). For the newer H200, **RunPod prices on-demand access at $4.39 per hour** (Source: www.runpod.io), while Microsoft's Azure AI Foundry Managed Compute tier lists **H100 access at a flat $8 per compute hour** for dedicated, fully managed GPU infrastructure (Source: azure.microsoft.com).

Multiplying those two variables out produces the real answer to "how much does it cost to run DeepSeek R1." A minimally viable **16x H100** cluster at CoreWeave's HGX rate of $4.76 per GPU per hour costs approximately **$76.16 per hour**, or roughly **$54,835 per month** run continuously (24 hours a day, 30 days), before accounting for the fact that idle capacity between requests is still billed. Against that fixed cost, the SemiAnalysis-benchmarked $1.365 to $15.692 per million tokens on H100 only becomes cheaper than buying tokens from a third-party host, such as DeepInfra which is confirmed the most affordable tracked provider at $0.56 per million tokens (Source: artificialanalysis.ai), once utilization is high and sustained enough to amortize that fixed hourly cost across a very large token volume. For comparison, DeepSeek's newer, non-reasoning V4 Pro model can be purchased through Together AI at **$1.74 per million input tokens and $0.20 per million cached input tokens**, with output at **$3.48 per million tokens** (Source: www.together.ai), a useful anchor for gauging how much of R1's self-hosting cost premium is attributable to the reasoning architecture specifically versus general model-serving overhead. This is precisely the dynamic that led one AWS-based case study, detailed later in this report, to conclude that self-hosting is not economical below roughly 100 concurrent internal users (Source: www.zenml.io).

Comparative Context and Market Positioning

Placed against the broader large language model (LLM) market, DeepSeek R1's cost structure, whether accessed via API or self-hosted, sits at the aggressive end of the pricing spectrum. It is worth remembering that R1 itself was built by applying reinforcement learning on top of the DeepSeek-V3 base checkpoint (Source: huggingface.co), so any self-hosting cost calculation for R1 inherits the same 671 billion parameter, 37 billion activated MoE footprint as V3 itself (Source: arxiv.org). OpenAI's current flagship reasoning-capable model prices at **$5.00 per million input tokens and $30.00 per million output tokens**, with cached input at **$0.50** (Source: openai.com), while its legacy **o1** model, R1's direct 2025-era competitor, remains listed at **$15.00 input and $60.00 output per million tokens** (Source: developers.openai.com) (Source: developers.openai.com). Even DeepSeek's own successor models undercut that historical o1 pricing substantially: current-generation **deepseek-v4-flash** costs **$0.14 input (cache miss) and $0.28 output per million tokens** (Source: api-docs.deepseek.com) (Source: api-docs.deepseek.com).

Table 1 below summarizes verified per-million-token pricing across the major hosted channels for DeepSeek R1, alongside comparable closed-model reference points, all captured as of July 2026.

PROVIDER / MODEL	INPUT $ / 1M TOKENS	OUTPUT $ / 1M TOKENS	NOTES
DeepSeek official API, R1 launch pricing	$0.55 (cache miss), $0.14 (cache hit)	$2.19	Original 2025 `deepseek-reasoner` pricing; alias retires July 24, 2026 (Source: api-docs.deepseek.com) (Source: api-docs.deepseek.com)
DeepInfra (R1 0528)	$0.50	$2.15	Cheapest tracked third-party host (Source: artificialanalysis.ai) (Source: artificialanalysis.ai)
NovitaAI (via OpenRouter)	$0.70	$2.50	Highest token-share provider on OpenRouter (Source: openrouter.ai)
Microsoft Azure AI Foundry (R1 Global)	$1.35	$5.40	Fully managed, pay-as-you-go (Source: azure.microsoft.com)
Together AI (R1 0528)	$3.06 (implied)	$3.40 (blended)	Most expensive tracked provider (Source: artificialanalysis.ai)
DeepSeek current API, deepseek-v4-flash	$0.14 (cache miss)	$0.28	Successor model now served by `deepseek-reasoner` alias (Source: api-docs.deepseek.com)
OpenAI o1 (official)	$15.00	$60.00	R1's direct closed-model competitor at launch (Source: developers.openai.com) (Source: developers.openai.com)
OpenAI current flagship (GPT-5.6 Sol)	$5.00	$30.00	Current-generation closed-model reference point (Source: openai.com)
Self-hosted, H100, benchmarked (SemiAnalysis)	N/A (blended)	$1.365 to $15.692	Varies with per-user interactivity target, excludes idle time and ops overhead (Source: inferencex.semianalysis.com)
Self-hosted, B200, benchmarked (SemiAnalysis)	N/A (blended)	$0.113 to $1.088	Same methodology, newer accelerator (Source: inferencex.semianalysis.com)

The pattern in *Table 1* is unambiguous: DeepSeek R1, whether purchased through DeepSeek's own historical API pricing or through third-party rehosting, undercuts equivalent closed reasoning models by roughly an order of magnitude at list price, and self-hosting on modern accelerators like B200 can undercut even that when utilization is high. But the table also shows that self-hosted cost is a range, not a fixed number, because it depends on the interactivity target chosen, whereas API pricing is a fixed quote regardless of how a buyer's traffic is shaped. This is the core trade-off a self-hosting decision has to weigh: predictable, published per-token pricing from a host, against variable, workload-dependent economics that can beat or lose to that host price depending entirely on operational discipline.

Data Analysis and Evidence

Quantifying DeepSeek's own economics helps calibrate expectations for what self-hosters are attempting to replicate. DeepSeek's technical report for the V3 base model, on which R1 was subsequently built through reinforcement learning, states that pre-training required **2.664 million H800 GPU hours**, context extension **119,000 GPU hours**, and post-training **5,000 GPU hours**, for a **total of 2.788 million GPU hours** (Source: arxiv.org). Assuming a rental price of **$2 per H800 GPU hour**, the paper states total training costs came to **only $5.576 million** (Source: arxiv.org), a figure that generated significant industry debate. SemiAnalysis's own analysis cautions that this number reflects only the final training run's compute cost and excludes research, ablation experiments, and prior infrastructure spend, estimating DeepSeek's actual total server capital expenditure at

approximately $1.6 billion, with a further **$944 million** in associated operating costs (Source: newsletter.semianalysis.com). The same analysis estimates DeepSeek and its parent hedge fund, High-Flyer, have access to roughly **10,000 H800 GPUs and about 10,000 H100 GPUs** (Source: newsletter.semianalysis.com), a compute base that dwarfs what any individual self-hosting team would deploy.

A key architectural reason DeepSeek can offer such low inference pricing is **Multi-head Latent Attention (MLA)**, an innovation SemiAnalysis credits with reducing the amount of key-value (KV) cache memory required per query by approximately **93.3%** versus standard attention mechanisms (Source: newsletter.semianalysis.com). Because KV cache size directly determines how many concurrent requests a fixed pool of GPU memory can serve, this reduction is a major driver of DeepSeek's favorable cost-per-token economics relative to attention-heavy architectures, and it is a factor any self-hoster inherits automatically simply by running the same architecture, regardless of which hardware vendor supplies the GPUs. The cost gap between DeepSeek's reasoning and non-reasoning model lines is itself informative for self-hosting decisions: Azure's own catalog prices the non-reasoning **DeepSeek V3 Global** at **$1.14 per million input tokens and $4.56 per million output tokens** (Source: azure.microsoft.com), meaningfully below R1's $1.35 and $5.40 on the same platform, and the newest **DeepSeek-V4 Flash Global** at just **$0.19 input and $0.51 output per million tokens** (Source: azure.microsoft.com), a gap attributable almost entirely to the additional inference-time compute R1 spends generating reasoning tokens before its final answer.

Independent throughput benchmarking illustrates how quickly the frontier moves. In March 2025, inference provider Avian.io announced achieving **303 tokens per second** running DeepSeek R1 on NVIDIA's then-new **Blackwell B200** architecture in collaboration with NVIDIA, describing it as "world leading inference performance" for the model (Source: www.reddit.com). Provider-level benchmarking from Artificial Analysis shows current-generation output speeds for the R1 0528 checkpoint varying from **164.9 tokens per second at Google Vertex**, the fastest tracked provider, down to **31.6 tokens per second at DeepInfra**, a **5.2x spread** attributable to differing hardware allocation and batching strategy rather than any change in the model itself (Source: artificialanalysis.ai). This spread matters directly for cost-per-token math, because a provider serving more tokens per second per GPU can spread the same fixed hourly hardware cost across a larger token volume, lowering the effective price it can profitably charge, exactly the lever self-hosters must also pull to make their own infrastructure investment pay off.

Case Studies and Real-World Examples

LiftOff LLC: Self-Hosting on AWS as a ChatGPT Plus Alternative

LiftOff LLC, a startup, conducted a documented evaluation of self-hosting DeepSeek R1 distilled models (1.5B, 7B, 8B, and 16B parameters) on **AWS EC2** GPU instances using Docker, Ollama, and OpenWebUI, explicitly testing the approach as a potential replacement for paid AI coding assistants (Source: www.zenml.io). The team found a single **g5g.2xlarge instance running continuously cost approximately $414 per month**, versus **$20 per user per month for ChatGPT Plus** (Source: www.zenml.io). Beyond the direct hardware bill, the team identified significant hidden costs, including setup time, DevOps overhead, and the engineering effort required for optimization and troubleshooting, that widened the economic gap further (Source: www.zenml.io). Their conclusion was direct: self-hosting only becomes economically sensible at much larger scale, or where privacy, customization, or regulatory requirements independently justify the added complexity (Source: www.zenml.io).

Perplexity AI: Sovereign Hosting for Data-Residency Reasons

When search engine Perplexity AI integrated DeepSeek R1 into its product in January 2025, it did not simply call DeepSeek's own API. Perplexity announced that R1 access through its platform is "hosted exclusively in US & EU data centers, your data never leaves Western servers," explicitly to isolate user queries from the Chinese vendor's infrastructure (Source: x.com). This is a case of self-hosting driven not primarily by per-token cost arbitrage but by data-sovereignty requirements, illustrating that the "self-hosting cost per million tokens" question is frequently inseparable from a parallel compliance question: the open MIT license is what made this sovereign re-hosting legally straightforward in the first place.

Italy's Garante and the Regulatory Case for Self-Hosting

On January 30, 2025, Italy's data protection authority, the **Garante**, ordered DeepSeek to block its consumer chatbot application in the country after the company failed to adequately address the regulator's questions about personal data handling, including whether user data was stored in China (Source: www.reuters.com). Regulators in Ireland and France opened parallel inquiries around the same period (Source: www.reuters.com). For European enterprises, this regulatory episode is frequently cited as the practical trigger for self-hosting R1's open weights entirely on domestic or approved infrastructure rather than routing any traffic through DeepSeek's hosted service, sidestepping the underlying data-residency dispute altogether by using the MIT-licensed weights independently of the vendor's own servers.

Microsoft Azure AI Foundry: Enterprise Managed Self-Hosting

Rather than leaving enterprises to build their own GPU clusters from scratch, Microsoft added DeepSeek R1 to its **Azure AI Foundry** model catalog, describing it as joining "a diverse portfolio of over 1,800 models" available for enterprise deployment (Source: techcommunity.microsoft.com). Azure's Managed Compute tier lets customers run R1 on dedicated **A100, H100, H200, or MI300 GPU** infrastructure without directly managing the underlying hardware, billed at **$4 per compute hour for A100 80G and $8 per compute hour for H100**, scaling automatically with traffic (Source: azure.microsoft.com) (Source: azure.microsoft.com). Enterprises requiring guaranteed capacity rather than pure pay-as-you-go billing can instead reserve Provisioned Throughput Units (PTUs), with **DeepSeek R1 Global** requiring a minimum of **100 PTUs priced at $1 per PTU-hour** (Source: azure.microsoft.com), a third pricing model that sits between fully self-managed infrastructure and pure metered API consumption: the enterprise pays a premium over raw GPU rental for managed provisioning, patching, and scaling, but retains more control over data flow and model version than it would through the pay-as-you-go tier alone.

A Community-Scale Self-Host Attempt (Hypothetical Example)

Consider a hypothetical mid-sized software company evaluating whether to self-host DeepSeek R1 for roughly 50 internal engineers with moderate, overlapping usage, a scenario closely mirroring a real discussion thread in the r/selfhosted community where a poster asked exactly this question with a $10,000 budget (Source: www.reddit.com). Community responses converged on the view that reasonable quality at that scale requires either the full model, quantized to roughly **IQ4_XS precision across 16 consumer GPUs such as RTX 3090s or 4090s**, at an estimated **$4,000 in non-GPU server costs alone before purchasing the GPUs themselves** (Source: www.reddit.com), or a smaller distilled 70B checkpoint requiring at least **40GB of VRAM**, achievable on two to three consumer-grade GPUs (Source: www.reddit.com). Another commenter in the same thread was blunter, replying simply that self-hosting at that scale was "probably not worth it" given the slow performance quantized full-size deployments deliver (Source: www.reddit.com). This scenario illustrates the practical middle ground many organizations actually occupy: too small to justify a data-center-grade H100 cluster, but large enough that consumer GPU workarounds introduce real reliability and quality trade-offs.

Implications and Future Directions

The trajectory of DeepSeek's own product roadmap suggests the self-hosting calculus will keep shifting under practitioners' feet. The imminent **July 24, 2026** retirement of the `deepseek-reasoner` API alias in favor of `deepseek-v4-flash` (Source: api-docs.deepseek.com) is itself evidence that vendors treat hosted model names as disposable, while the underlying open-weight checkpoints, protected by the MIT license, persist independently of any single company's roadmap decisions (Source: huggingface.co). Organizations with compliance or reproducibility requirements around a specific model version, rather than "whatever DeepSeek's API currently serves," have a structural reason to prefer self-hosting regardless of near-term cost, because it is the only way to guarantee model behavior does not silently change underneath them.

Hardware economics are also moving quickly in self-hosters' favor. The SemiAnalysis benchmarking data shows **NVIDIA's B200** delivering roughly **12x lower cost per million tokens than H100** at equivalent interactivity settings (Source: inferencex.semianalysis.com), an advantage rooted in native FP8 tensor-core support that older architectures like A100 simply lack (Source: www.theriseunion.com). NVIDIA's own H200 specification sheet confirms nearly double the memory capacity of H100 (141GB versus 80GB) at meaningfully higher bandwidth (**4.8 terabytes per second**) (Source: www.nvidia.com), directly reducing the GPU count needed for the full 671B model. As newer accelerator generations reach cloud rental markets at competitive hourly rates, the utilization threshold at which self-hosting beats API pricing should continue to fall, though it will likely never reach zero given that idle GPU-hours remain a fixed cost that API consumption avoids entirely.

Regulatory pressure represents a second, independent force reshaping this decision. Following Italy's block and parallel inquiries from Ireland and France (Source: www.reuters.com), a growing number of European and regulated-industry organizations appear willing to accept a self-hosting cost premium specifically to avoid any data transit through DeepSeek's own infrastructure, a pattern already visible in Perplexity's sovereign-hosting approach (Source: x.com). For such organizations, the relevant comparison is not self-hosting versus DeepSeek's own API at all, but self-hosting versus a compliant third-party rehoster such as Azure, meaning the effective "cost of self-hosting" should be benchmarked against Azure's **$1.35 to $5.40 per million token** managed pricing (Source: azure.microsoft.com) rather than DeepSeek's own historical rates.

Frequently Asked Questions (FAQs)

What is DeepSeek R1's API pricing per million tokens? At launch, DeepSeek priced R1 access at **$0.55 per million input tokens on a cache miss, $0.14 per million input tokens on a cache hit, and $2.19 per million output tokens** (Source: api-docs.deepseek.com) (Source: api-docs.deepseek.com). That specific `deepseek-reasoner` alias is scheduled for deprecation on **July 24, 2026**, after which it will route to `deepseek-v4-flash` pricing instead (Source: api-docs.deepseek.com).

What GPU hardware does self-hosting DeepSeek R1 require? Vendor guidance recommends **8x NVIDIA H200 or H20 (141GB each)** or **16x H100 or H800 (80GB each)** for native FP8 inference, or **16x A100 or A800 (80GB each)** if BF16 conversion is required (Source: theriseunion.com). Smaller distilled models, from 1.5B to 70B parameters, can run on single consumer or professional GPUs instead (Source: huggingface.co).

What is a typical cost per million tokens for a self-hosted LLM (large language model) generally? For DeepSeek R1 specifically, benchmarked self-hosted costs on H100 range from **$1.365 to $15.692 per million tokens** depending on batching and latency targets, versus **$0.113 to $1.088** on the newer B200 accelerator (Source: inferencex.semianalysis.com).

Is there a reliable self-hosted LLM cost calculator? No single authoritative calculator exists; the underlying variables (GPU hourly rate, batching efficiency, and target latency) vary too much by deployment to reduce to a single number. Aggregators such as OpenRouter report weighted average prices that shift over time as providers adjust capacity (Source: openrouter.ai), underscoring why a static calculator cannot substitute for benchmarking a specific workload against the ranges this report documents.

How does DeepSeek R1 compare to GPT-4 on cost? DeepSeek R1's launch API pricing of **$0.55 input and $2.19 output per million tokens** (Source: api-docs.deepseek.com) undercut OpenAI's contemporaneous reasoning model o1, officially priced at **$15.00 input and $60.00 output** (Source: developers.openai.com), by well over an order of magnitude, and continues to undercut OpenAI's current flagship pricing of **$5.00 input and $30.00 output** (Source: openai.com).

Is DeepSeek R1 genuinely open source, and can it be used commercially? Yes. Both the model weights and code repository are released under the **MIT License**, which explicitly permits commercial use, modification, and distillation into other models (Source: huggingface.co). Note that the smaller distilled variants inherit separate licenses (Apache 2.0 for the Qwen-based distills, Meta's Llama license for the Llama-based distills) from their respective base models (Source: huggingface.co).

Does self-hosting make sense for a small team? Documented evidence suggests no, below a certain scale. LiftOff LLC's case study found self-hosting distilled R1 models cost roughly **20x more per user per month** than ChatGPT Plus at under-100-user scale, once instance costs alone were compared (Source: www.zenml.io), a conclusion echoed by community discussion threads weighing similar trade-offs at a comparable scale (Source: www.reddit.com).

Conclusion

Self-hosting DeepSeek R1 is economically justifiable at scale and for specific compliance or reproducibility needs, but it is not automatically cheaper than buying tokens from an API, and the evidence gathered in this report supports that conclusion in both directions. On the hardware side, running the full 671B parameter model requires an 8 to 16 GPU cluster of data-center accelerators, a fixed cost running into tens of thousands of dollars per month regardless of actual usage, whose benchmarked cost per million tokens ranges from roughly $0.11 on the newest B200 hardware at high utilization to nearly $16 on older H100 hardware at low-latency, low-batch settings. On the API side, DeepSeek's own historical pricing and the wider third-party hosting market, spanning DeepInfra, Azure, Together AI, and others, already deliver R1-class reasoning at prices an order of magnitude below closed-model competitors, without any infrastructure burden at all.

The decision therefore comes down to three factors this report has documented directly: sustained utilization (self-hosting wins only at high, continuous load), data-residency or reproducibility requirements (self-hosting wins regardless of raw cost when compliance mandates it, as with Perplexity's sovereign hosting and the regulatory pressure following Italy's Garante action), and organizational scale (documented case evidence puts the break-even point well above the casual or small-team usage level). The imminent retirement of DeepSeek's own `deepseek-reasoner` API alias on July 24, 2026 adds a final, time-sensitive consideration: teams that need the original R1 model's exact behavior indefinitely now have a concrete reason to secure their own copy of the MIT-licensed weights rather than assume any hosted API will keep serving that exact model going forward. For most teams evaluating this question today, the practical answer is to start with a third-party hosted API, benchmark actual token volume and latency needs against the cost curves this report has laid out, and only invest in dedicated GPU infrastructure once usage data shows the crossover point has genuinely been reached.

Tags: deepseek r1, self-hosting cost, llm pricing, gpu inference cost, deepseek api pricing, cost per million tokens, open source llm, gpu requirements, inference benchmarking

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will GPUSmith or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.