AI INFRASTRUCTURE · AUGUST 12, 2026

LLM Inference API Providers in August 2026: Pricing, Speed and Reliability Compared

Pricing and reliability comparison of inference API providers as of Aug 2026: first-party frontier APIs (OpenAI, Anthropic, Google, DeepSeek), open-weight hosts (Groq, Cerebras, Together, DeepInfra, Fireworks) and aggregators (OpenRouter), with cache/batch discounts, measured latency and uptime data.

7 min read · 17 sources

2026 pricing | API pricing | LLM APIs | inference providers | reliability | uptime

CONTENTS

First-party frontier APIs: where the flagship price is set 1
Open-weight hosts: speed versus value 2
Reliability: the dimension that separates the market 4
What changed in 2026 5
Bottom line 6
Sources 6

Buying model access in August 2026 means choosing between three very different kinds of vendor: first-party frontier APIs (OpenAI, Anthropic, Google, DeepSeek), serverless open-weight hosts (Groq, Cerebras, Together AI, DeepInfra, Fireworks), and aggregator/gateway layers (OpenRouter and its peers). The same model family can cost 10–600× differently depending on which provider serves it, and reliability — not headline price — is where the providers diverge most sharply. All prices below are USD per 1M tokens, checked against official pricing pages in the week of 11 August 2026.

$0.14/$0.28	$5/$30	~120 ms	99.42%	5.5%
DeepSeek V4 Flash (in/out)	GPT-5.6 Sol flagship (in/out)	Cerebras p50 time-to-first-token	Claude API measured 30-day uptime	OpenRouter card credit-purchase fee
cheapest frontier-tier API		vs ~820 ms OpenAI	no published SLA	

First-party frontier APIs: where the flagship price is set

The big-four first-party APIs set the ceiling and the floor of the market. OpenAI's current lineup is the GPT-5.6 family, GA since 9 July 2026: **Sol at $5/$30, Terra at $2/$12, Luna at $0.20/$1.20** (input/output per 1M) [1] [10]. Cached input bills at 10% of standard rates, the Batch and Flex APIs halve everything, and "Fast mode" (renamed from Priority on 30 July 2026) doubles standard pricing;

long-context calls over ~272K tokens run 2× input and 1.5× output. Data-residency endpoints add a 10% uplift for models released on or after 5 March 2026[1]. On 30 July 2026 OpenAI cut Terra 20% and Luna 80% while Sol held its launch price[10].

Anthropic prices Claude **Opus 5 at $5/$25, Sonnet 5 at $2/$10, Haiku 4.5 at $1/$5**, and the agent-oriented Fable 5 at $10/$50[2]. Two things worth knowing: the $2/$10 Sonnet 5 rate, announced as introductory pricing through 31 August, is now standard — the scheduled 1 September rise to $3/$15 **will not occur**[2]. And models from 4.7 on use a tokenizer that produces roughly 30% more tokens for the same text, silently inflating bills versus older models[2]. Cache hits bill at $0.50/$0.20/$0.10 for Opus/Sonnet/Haiku; cache writes at 1.25× (5-minute) or 2× (1-hour). Bedrock and Vertex regional endpoints carry a 10% premium[2].

Google's Gemini API is cheaper per tier: **Gemini 3.6 Flash at $1.50/$7.50** (cache $0.15), 3.5 Flash at $1.50/$9.00, and **3.5 Flash-Lite at $0.30/$2.50** — the cheapest first-party GA model with a serious quality score[3]. Batch halves all rates; a "Priority" tier runs ~1.8× standard[3]. DeepSeek undercuts everyone: **V4 Flash at $0.14/$0.28 with cache hits at $0.0028**, and V4 Pro at $0.435/$0.87 — the cheapest frontier-adjacent APIs on the market, both with 1M-token context[4]. Caveat: DeepSeek officially warns it plans a "significant" price increase in the near future[4].

PROVIDER	MODEL	INPUT	CACHED IN	OUTPUT	BATCH	CONTEXT
OpenAI	GPT-5.6 Sol	$5.00	$0.50	$30.00	50% off	1.05M
OpenAI	GPT-5.6 Terra	$2.00	$0.20	$12.00	50% off	1.05M
OpenAI	GPT-5.6 Luna	$0.20	$0.02	$1.20	50% off	1.05M
Anthropic	Claude Opus 5	$5.00	$0.50	$25.00	—	1M
Anthropic	Claude Sonnet 5	$2.00	$0.20	$10.00	—	1M
Anthropic	Claude Haiku 4.5	$1.00	$0.10	$5.00	—	1M
Google	Gemini 3.6 Flash	$1.50	$0.15	$7.50	50% off	—
Google	Gemini 3.5 Flash-Lite	$0.30	—	$2.50	50% off	—
DeepSeek	V4 Flash	$0.14	$0.0028	$0.28	—	1M
DeepSeek	V4 Pro	$0.435	$0.0036	$0.87	—	1M

Open-weight hosts: speed versus value

Once you pick an open-weight model, the host decides the price and the latency. Groq and Cerebras sell speed from purpose-built silicon; Together, DeepInfra and Fireworks sell breadth and low cost from GPU fleets[16].

On Groq's LPU, Llama 3.1 8B Instant runs **$0.05/$0.08 at ~560 tokens/s**, GPT-OSS 20B $0.075/$0.30 at ~1,000 tok/s, GPT-OSS 120B $0.15/$0.60 at ~500 tok/s, and Llama 3.3 70B $0.59/$0.79 at ~280 tok/s [5]. Batch is 50% off and cached input 50% off — a smaller cache discount than the first-party providers [16]. The catalog is the catch: six self-serve models, no DeepSeek, no fine-tuning outside enterprise [5] [16]. Cerebras's wafer-scale engine is the throughput king — 2,100–2,500 tok/s on Llama 3.3 70B, roughly 3–8× Groq on independent numbers — but prices about 40–70% higher: **Llama 3.3 70B at ~$0.85/$1.20 and GPT-OSS 120B at ~$0.35/$0.75** [16] [17]. Its free tier gives $5 of credits and a Developer tier starts at $10 [6].

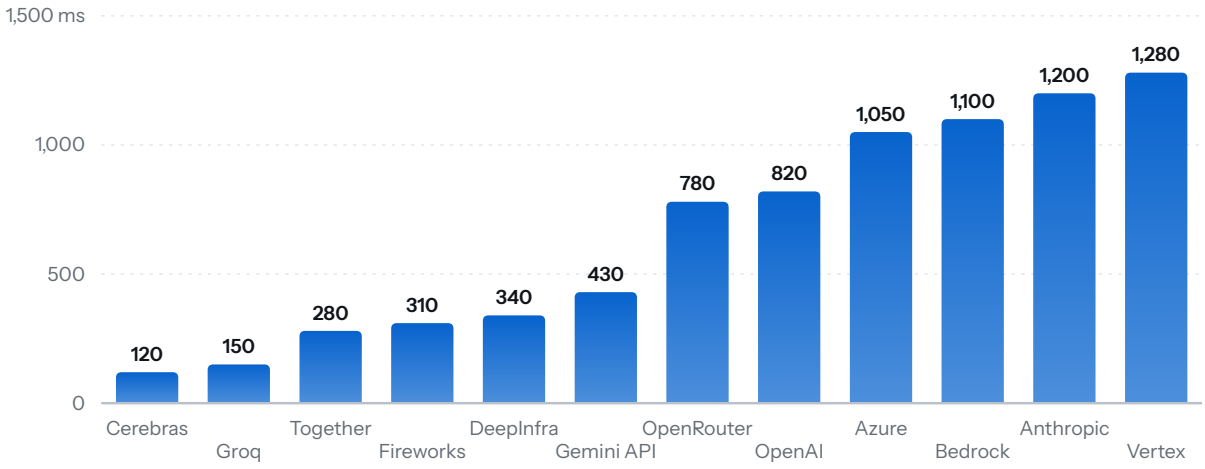
	Groq (LPU)	vs	Cerebras (WSE-3)
Llama 3.3 70B in/out	$0.59 / $0.79	·	~$0.85 / $1.20
Llama 3.3 70B throughput	~294–750 tok/s	·	~2,100–2,500 tok/s
GPT-OSS 120B in/out	$0.15 / $0.60	·	~$0.35 / $0.75
Self-serve catalog	6 models	·	8 models

Prices checked 11 Aug 2026 [16]; throughput spans vendor and independent (Artificial Analysis) measurements.

Together and DeepInfra are the value picks for open weights. Together serves **DeepSeek V4 Flash at $0.14/$0.28 — the same as DeepSeek first-party** — plus Kimi K3 at $3/$15, GLM-5.2 at $1.40/$4.40 and MiniMax M3 at $0.30/$1.20, with the broadest model catalog and self-serve fine-tuning [7]. DeepInfra is often cheapest of all: **DeepSeek V4 Flash (0731) at $0.08/$0.18, V4 Pro at $1.30/$2.60, Kimi K3 at $2.85/$14.25** [8]. Fireworks targets production serving (function calling, structured output) and publishes GPU on-demand rates from $7/hr (H100/H200) to $18/hr (GB300) rather than per-token lists on its pricing page; region-restricted deployments run 1.5× [9].

Measured time-to-first-token (independent April 2026 harness, 1,000 calls/provider) shows the speed gap is architectural, not tuning: **Cerebras ~120 ms and Groq ~150 ms p50, versus ~820 ms OpenAI, ~1,050 ms Azure OpenAI, ~1,100 ms Bedrock, ~1,200 ms Anthropic, ~1,280 ms Vertex** [12].

Median time-to-first-token, p50 (ms, lower is better)



VerticalAPI harness, April 2026; models since superseded [12].

CAUTION

The latency numbers above were measured in April 2026 on the model lineup of that moment, and prices on every provider "move without announcements" [12]. Re-verify both against the provider's own page before budgeting.

Reliability: the dimension that separates the market

Reliability is where the marketing pages go quiet. A February 2026 measurement round found the first-party APIs substantially less available than the cloud infrastructure that hosts them [13]:

PROVIDER / ROUTE	MEASURED UPTIME (30-DAY)	PUBLISHED SLA
Claude API (Anthropic)	99.42%	None
Claude.ai	99.32%	None
OpenAI API	~99.5%	99.9%, Enterprise tier only
Azure OpenAI	~99.9%	99.9%
Google Vertex AI	~99.9%	99.9%
Amazon Bedrock	~99.9%	99.9%

99.42% sounds fine until you convert it: nearly **5 hours of downtime per month**, and LLM outages are global rather than region-scoped — there is no geographic failover you can buy from the same vendor [13]. The cloud resellers (Azure, Bedrock, Vertex) carry real SLAs and measured ~99.9%, but add latency and, on Anthropic's regional endpoints, a 10% price premium [2] [12] [13].

OpenRouter is the most convenient aggregator — 400+ models behind one OpenAI-compatible API, with automatic provider failover on by default and model fallbacks you configure [15]. But it is not a reliability product: **no SLA, three documented outages in eight months (50, 38 and 35 minutes), a 5.5% credit-card purchase fee (5.0% crypto) with credits expiring after 365 days, and 25–40 ms added latency** [14]. One February 2026 incident even returned misleading 401 "User not found" errors during infrastructure failure (since fixed to 503) [14]. The flip side: OpenRouter's zero-completion insurance means you don't pay for requests that ultimately fail, and provider failover means worst-case uptime beats any single direct integration [15]. Its pass-through token prices carry no markup on the tokens themselves — the fee lives on credit loading instead [14].

KEY FINDING

The practical reliability pattern in 2026 is multi-provider: route traffic through an aggregator or gateway with failover, keep a cheaper model (Gemini Flash-Lite, DeepSeek, Groq Llama) as the fallback tier, and never assume a first-party API's uptime is SLA-backed [13] [15].

What changed in 2026

28 Aug 2025 — **OpenRouter database outage**
~50 min downtime; first of three incidents in 8 months [14]

9 Jul 2026 — **GPT-5.6 family GA**
Sol $5/$30, Terra $2/$12, Luna $0.20/$1.20, 1.05M context [10]

17–19 Feb 2026 — **OpenRouter cache-layer outages**
38 min and 35 min; 401 errors during failure [14]

30 Jul 2026 — **OpenAI price cuts**
Terra −20%, Luna −80%; Priority renamed Fast mode [1] [10]

Aug 2026 — **Anthropic holds Sonnet 5 at $2/$10**
Sept 1 increase to $3/$15 cancelled; intro price made standard [2]

Announced — **DeepSeek price rise planned**
'Significant' increase expected; plan usage accordingly [4]

Two structural trends are visible through 2026: OpenAI and Anthropic keep a wide price ladder so flagship cost is optional (GPT-5.5 Pro still lists at $30/$180; o1-pro at $150/$600 — a 600× spread to GPT-5 nano at $0.05/$0.40) [1] [10], while DeepSeek and the GPU hosts keep compressing the commodity floor toward $0.10–0.30 per 1M input [4] [7] [8]. Meanwhile the speed leaders — Groq, Cerebras — compete on TTFT that is 5–8× faster than the first-party APIs, at the price of a six-to-eight-model catalog with no fine-tuning path [16].

Bottom line

For a production system in August 2026: use **Claude Sonnet 5 or GPT-5.6 Terra as the quality tier** ($2/$10 and $2/$12), **Gemini 3.5 Flash-Lite or DeepSeek V4 Flash as the volume tier** ($0.30/$2.50 and $0.14/$0.28), and **Groq/Cerebras for latency-critical streaming** (sub-200 ms TTFT on Llama-class models). Take cache-hit discounts only where your prompts actually reuse prefixes, use Batch/Flex for anything non-interactive, and treat OpenRouter as routing convenience with a 5.5% credit fee and no SLA — pair it with direct access for the models you depend on. The single most expensive mistake is defaulting every call to a flagship: matching model to task is worth more than any provider discount [10] [12].

SOURCES

1 OpenAI API Pricing (official docs)
https://developers.openai.com/api/docs/pricing

2 Pricing — Claude Platform Docs (official)
https://platform.claude.com/docs/en/about-claude/pricing

3 Gemini Developer API pricing (official)
https://ai.google.dev/gemini-api/docs/pricing

4 Models & Pricing — DeepSeek API Docs (official)
https://api-docs.deepseek.com/quick_start/pricing

5 Supported Models — GroqDocs (official)
https://console.groq.com/docs/models

6 Cerebras Pricing (official)
https://www.cerebras.ai/pricing

7 Together AI Pricing (official)
https://www.together.ai/pricing

8 DeepInfra Pricing (official)
https://deepinfra.com/pricing

9 Fireworks AI Pricing (official)
https://fireworks.ai/pricing

10 OpenAI API Pricing (August 2026) — BenchLM.ai
https://benchlm.ai/openai/api-pricing

11 LLM API Pricing Comparison & Calculator — BenchLM.ai
https://benchlm.ai/llm-pricing

12 LLM Benchmark 2026: latency and cost, 26 providers — VerticalAPI
https://verticalapi.com/benchmark/

13 AI uptime SLA: why your business needs a multi-model fallback strategy — Universal.cloud
https://universal.cloud/en/blog/ai-uptime-vergeten-risico/

14 Is OpenRouter Reliable? An Honest Review for Production Use (2026) — ofox.ai
https://ofox.ai/blog/is-openrouter-reliable-honest-review-2026/

15 OpenRouter Reliability & Automatic Failover — OpenRouter Blog
https://openrouter.ai/blog/insights/reliability-failover/

16 Groq vs Cerebras: Cheapest Inference Chip to Rent in 2026 — Spheron
https://www.spheron.network/blog/groq-lpu-vs-cerebras-wse-cheapest-inference-2026/

17 genai-prices: Cerebras provider pricing — pydantic (GitHub)
https://github.com/pydantic/genai-prices/blob/main/prices/providers/cerebras.yml

Researched and written with deepseek/deepseek-v4-flash-0731, from the sources listed above. Figures are as reported by those sources. Exported November 29, 2026 from https://intel.innerpulse.net/doc/llm-inference-api-providers-in-august-2026-pricing-speed-and-reliability, where the current version lives.