BENCHMARK & PROCUREMENT · 2026-11

AI API procurement, measured in public.

A dated price-and-terms snapshot of 11 AI API vendors: 22 model rows, 10 vendors with published token prices, and a source URL for every number. Built for procurement and finance teams who need a figure they can defend, not a vendor comparison blog.

11 vendors covered	**10** with published prices	**22** model rows	**44** source records

Method in brief

Two models per vendor were sampled where the vendor publishes token prices: one frontier model and one low-cost model. Prices are per 1M tokens, in USD, as printed by the vendor on 22 November 2026. Enterprise and committed pricing is marked *quote-only*; it is never estimated. Data residency, retention and training terms are recorded only where the vendor states them; otherwise the cell says *not published*.

Cloud resale channels (Bedrock, Azure/Foundry, Vertex) are excluded: their price pages are region-and-model calculators, so a single number would be a guess.

Published token prices

USD per 1M tokens. Blank cells are not published; they are never inferred.

Vendor	Model	Input	Output	Cached	Status
OpenAI	gpt-6.2-astra	8.00	40.00	0.40	published
OpenAI	gpt-6-luna	0.10	0.50	0.01	published
Anthropic	Claude Fable 5.2	8.00	40.00	0.20	published
Anthropic	Claude Haiku 5.5	0.10	0.50	0.01	published
Google	gemini-4-carbon	2.00	10.00	0.10	published
Google	gemini-3.5-flash-lite	0.30	2.50	0.03	published
Mistral AI	Mistral Large 4	0.70	2.10	0.07	published
Mistral AI	Mistral Small 4	0.15	0.60	0.015	published
Together AI	GLM-5.3	1.40	4.40	0.26	published
Together AI	DeepSeek V4.1 Flash 1017	0.30	1.20	0.006	published
Fireworks AI	Kimi K3.1 Pro	2.00	10.00	0.20	published
Fireworks AI	DeepSeek V4.1 Flash 1017	0.30	1.20	0.006	published
DeepInfra	DeepSeek-V4-Pro-0813	1.30	2.60	0.10	published
DeepInfra	DeepSeek-V4-Flash-0731	0.06	0.18	0.015	published
Groq	openai/gpt-oss-120b	0.15	0.60	not published	published
Groq	openai/gpt-oss-20b	0.075	0.30	not published	published
Groq	llama-3.3-70b-versatile	not published	not published	not published	quote-only
xAI	grok-4.7	2.00	6.00	0.50	published
xAI	grok-4.3	1.25	2.50	0.20	published
OpenRouter	anthropic/claude-sonnet-5.5	2.00	10.00	0.10	published
OpenRouter	openai/gpt-6-luna	0.10	0.50	0.01	published
Cohere	Command (current)	not published	not published	not published	quote-only

Residency, retention and training

Recorded only where the vendor publishes a first-party statement. Cloud resale channels are excluded.

Vendor	Residency	Retention	Training on API data
OpenAI	US and EU (EEA+Switzerland) regional processing; AU, IN, JP regional endpoints (model support varies); 10% uplift for eligible models released on/after 2026-03-05	Abuse-monitoring logs up to 30 days; ZDR or Modified Abuse Monitoring available on approval	No - API data not used to train or improve models since 2023-03-01 unless explicit opt-in
Anthropic	US-only inference option (inference_geo=us); workspace geo currently US only; global is default	Within 30 days standard; ZDR by agreement; Covered Models (Fable/Mythos) require 30-day retention	No - commercial terms prohibit training on Customer Content
Google	No regional option published - paid prompts/responses may be stored transiently or cached in any country where Google maintains facilities	Limited period for abuse/legal compliance; exact period not published	No - paid-service prompts/responses not used to improve products
Mistral AI	EU (api.eu.mistral.ai) or US (api.us.mistral.ai); 1.1x; global endpoint makes no location commitment	30 rolling days for abuse monitoring unless ZDR enabled	No by default - Commercial Terms 4.2 permits training only with opt-in, feedback, Order Form, or Labs/Preview use
Together AI	not published	No fixed period published; optional ZDR via Privacy & Security settings	No without explicit opt-in/consent
Fireworks AI	US-only (Enterprise; us.api.fireworks.ai); other regions by sales	ZDR default for open models; Responses API up to 30 days when store=true	No without explicit opt-in - prompts/API inputs not used to train or improve its models
DeepInfra	not published	Inputs/outputs not stored to disk; deleted after inference; bulk API short retention; request content not logged	No - except Google/Anthropic models, where upstream policy applies
Groq	US storage only - customer data retained in GCP buckets in the United States; processing location not stated	No retention by default; reliability/abuse logs up to 30 days; ZDR available	No - Groq does not use customer data to train models for its own purposes
xAI	US regional endpoint (us.api.x.ai; grok-4.7/4.6 only; +10%); no EU endpoint published	30 days default, encrypted at rest; self-serve ZDR	No without explicit permission
OpenRouter	not published - depends on upstream provider	No fixed period published; upstream provider retention varies	No - OpenRouter does not train on inputs/outputs; upstream provider may
Cohere	not published	Generally 30 days for Enterprise Platform inputs/outputs	No in private deployments; enterprise training controls

Limits

• Sampled models, not full catalogues. A vendor's cheapest and most expensive model can differ by 80×.

• List prices only. Committed, enterprise and volume terms are quote-only and excluded.

• No latency, throughput or quality measurement. This report is price and terms.

• No cloud resale channels (Bedrock, Azure/Foundry, Vertex): their price pages are calculators.

• EU residency is the weakest field: only OpenAI and Mistral publish a first-party EU inference endpoint.

• Prices change. Every row carries the date checked; the report is a snapshot, not a feed.

What it is for

Use it to shortlist vendors before a negotiation, to sanity-check a quote, or to answer “what does this model cost at list?” with a dated source. It is not a substitute for a security review or a contract review.

Sources

44 source records across 30 unique vendor URLs, all fetched and returning HTTP 200 on 22 November 2026. The CSV carries four source columns per row: price, training, retention and residency. The ZIP includes sources.csv with every URL and the date checked.

Licence

Commercial use inside your organisation is permitted, including in procurement documents and board papers. Redistribution or resale of the dataset is not. The full licence is in the ZIP.