LLM Token Pricing Comparison

Take to the boardroom · November 2026 · Executive Mind, Brisbane

Built from the beach, for the boardroom.

Reference card for business readers: token pricing across reasoning-capable AI channels — the names you know, the credible open-weight challengers, and the one-subscription bundler (Ollama Cloud) hosting 20+ of them under zero-retention terms. All rates USD per 1 million tokens unless noted. Sources and retrieval dates on every block.

Part 1 — Direct API Pricing

Verify against the provider's own pricing page before contracting. Batch = 50% off across OpenAI / Anthropic / Google / Bedrock.

Anthropic (Claude)

platform.claude.com/docs/en/about-claude/pricing · retrieved 2026-11-14 · [official]

MODEL	IN /1M	CACHED	OUT /1M	NOTES
Claude Fable 5.2 (flagship)	$8.00	$0.20	$40.00	Current flagship tier
Claude Opus 5.5	$4.00	$0.20	$20.00	
Claude Sonnet 5.5	$2.00	$0.10	$10.00	
Claude Haiku 5.5	$0.10–varies	$0.01	$0.50	Two-tier by prompt length
Claude Opus 4.5	$5.00	~$0.50	$25.00	200K ctx
Claude Sonnet 4.6 / 4.5	$3.00	$0.30	$15.00	200K ctx

DeepSeek

deepseek.com / api-docs.deepseek.com · retrieved 2026-11-14 · [official]

MODEL	OFF-PEAK IN	PEAK IN	OFF-PEAK OUT	PEAK OUT
deepseek-v4.1-flash	$0.15	$0.30	$0.60	$1.20
deepseek-v4-pro	$0.66	$1.32	$1.98	$3.96

Peak = 01:00–04:00 & 06:00–10:00 UTC weekdays (excl. CN holidays). Off-peak = half. 1M ctx.

Google Gemini

aistudio.google.com/docs/pricing · retrieved 2026-11-14 · [official + secondary cross-checks]

MODEL	IN /1M	CACHED	OUT /1M	NOTES
Gemini 3.1 Pro (promo)	$2.00 / $4.00	$0.20→$0.40	$10→$20	Promo ends 2026-11-30; price doubles Dec 1
Gemini 3 Flash (preview)	$0.50	$0.05	$3.00	1M ctx
Gemini 2.5 Flash-Lite	$0.10	$0.01	$0.40	Cheapest Gemini

Mistral

console.mistral.ai + truefoundry.com + eesel.ai + openrouter.ai + requesty.ai · retrieved 2026-11-14 · [secondary — verify preview vs list status before quoting]

MODEL	IN /1M	CACHED	OUT /1M	NOTES
Mistral Large 4 (preview)	$0.68	$0.07	$2.09	No published end date
Mistral Large 4 (list)	$1.36	$0.14	$4.18	Base for EU/US ×1.1, Priority ×1.75
Mistral Large 3 (prev)	$0.50	$0.05	$1.50	Still listed

Open-weight, ~1.05T-param MoE. Preview launched 2026-10-06. 1M ctx. **Tied-first on AA Cyber Index** — the security/enterprise pick. Sovereign/EU deployment, on-prem, audit-friendly — the natural Anthropic/OpenAI alternative for regulated AU workloads. **Runs Garrison, our Security & Treasury agent, via `mistral-large-4:cloud`.**

Direct API — continued

Moonshot AI (Kimi)

kimi.ai / platform.moonshot.cn via cloudzero + requesty · retrieved 2026-11-14 · [secondary]

MODEL	IN /1M	CACHED	OUT /1M	NOTES
Kimi K3.1 Pro (new Nov 2026)	$2.00	n/p	$10.00	1M ctx; Highspeed 2×
Kimi K3 (open weights)	$3.00	$0.30	$15.00	2.8T-param (Jul 2026)
Kimi K2.7 Code	$0.95	$0.19	$4.00	256K ctx

OpenAI

developers.openai.com/api/docs/pricing · retrieved 2026-11-14 · [official]

MODEL	IN /1M	CACHED	OUT /1M	NOTES
GPT-6 Astra (flagship)	$10.00	$1.00	$50.00	Batch $5/$25
GPT-6.1 Sol	$2.00	$0.10	$10.00	Fast $4/$20
GPT-6,1 Terra	$1.00	$0.05	$5:00	Fast $2/$10
GPT-6 Luna	$0.10	$0.01	$0.50	Fast $0.20/$1.00
GPT-5.6 Sol (promo)	$4.00	$0.40	$20.00	Promo until 2026-11-30

Z.ai (Zhipu) GLM

open.bigmodel.cn/pricing + z.ai · retrieved 2026-11-14 · [official CN + secondary intl]

MODEL	IN /1M	CACHED	Out /1M	NOTES
GLM-5.3 (open weights)	$1.40	$0.26	$4.40	1M ctx
GLM-5.3-Flash	$0.15	$0.03	$0.50	1M ctx
GLM-5.3-FlashX	$0.37	$0.075	$1.25	Faster variant

Other Notables

platform.minimax.io · docs.x.ai · alibabacloud.com · aws.amazon.com/bedrock/pricing · thinkingmachines.ai · openrouter.ai · retrieved 2026-11-14

MODEL	IN /1M	CACHED	OUT /1M	NOTES
MiniMax-M3 (≤512K) (50% off)	$0.30	$0.06	$1.20	Permanent launch discount
Grok 4.7 (flagship)	$2.00	$0.50	$6.00	500K ctx
Qwen3.8-Max (flagship)	$2.00	$0.25	$6.00	1M ctx
Nova Premier	$2.50	$0.625	$12.50	Amazon's model
Nova 2 Lite (preview)	$0.30	~$0.075	$2.50	Amazon's model; reasoning low/med/high
Inkling-2-small (Tinker)	$0.30	$0.03	$1.00	1M ctx; fine-tuning $1.50/M–$3.00/M
Muse Spark Open (hosted)	$1.25	$0.15	$4.25	1M ctx; uniform across Together / OpenRouter / Baseten

Inkling 2 Small — Thinking Machines Lab (Mira Murati), 2026-11-05. Sparse MoE 594B/17B. **#1 among open weights on AA-AnalystAgent and Dayjob:Finance**; Vals Finance 61.5%. Reasoning depth over speed. **Runs Kimi, our CFO, via `inkling-2-small:cloud`. Muse Spark Open** — Meta, under the **Muse Model License** (not OSI open source). AA Intelligence Index 39, DiligenceBench 59.0% (highest). Priced like a closed model — the license is why. **Runs Aria, our Web/SEO lead, via `muse-spark-open:cloud`.**

These figures change constantly — and not a month goes by without a new model. If you would like an updated version, email us anytime 24/7 at admin@executivemind.io. Request **"most current"** — we'll update this regularly — or **"as of this moment"** and we'll re-run the research on demand, so you never walk into a meeting with an outdated price list. And if a current model (Microsoft, Cohere, AI21 Labs, etc.) isn't on this list and you'd like to know where it stands — we'll let you know where it stands, too. **Just ask.**

Part 2 — Ollama Cloud: The Bundler Story

ollama.com/pricing · retrieved 2026-11-14

One subscription, 20+ open-weight models, zero data retention. **You are buying convenience, one bill, and privacy posture — not a discount on tokens.** Steady high-volume work favours direct API on unit price; bursty multi-model teams get one invoice and no key sprawl.

What Ollama actually sells

PLAN	PRICE/MO	CREDITS/MO	CONCURRENCY	NOTES
Pay as you go	$0	buy credits	3	No service fee; US-hosted (EU overflow)
Pro	$20	$60	3	+ larger pro models, Fast mode coming
Max	$100	$300	10	+ early access to newest models
Team (early access)	$500	$1,000 shared	10	Unlimited seats, centralised billing/admin
Enterprise	custom	custom	custom	SSO, model access controls, budgets per user/key

Credits refresh monthly, never roll over. US-hosted (Europe overflow) under NVIDIA Cloud Partners — no-log / no-train / **zero-data-retention**.

When Ollama Cloud wins vs paying providers direct

WINS when

- One invoice, one API key, instant access to 20+ open-weight frontier models.
- Data-handling posture matters: zero-retention, no-training across all hosted models — a single contractual answer instead of six vendor DPAs.
- Bursty/unpredictable load; concurrency 3–10 with monthly credit pooling.
- You want open weights with the option to move in-house later.

LOSES when

- You need frontier closed models (Claude, GPT-6, Gemini 3 Pro) — Ollama does not host them.
- Sustained high-volume single-model workloads: direct API has no concurrency ceiling and no $500/mo floor.
- You need SLA-backed uptime, AU data residency, or audit certifications.
- Latency-critical paths: shared queues beyond your concurrency cap.



Compiled by Aria
Web, SEO & Brand
Executive Mind · Brisbane, Queensland
Designed by Aria.

CONTACT & SOURCES

admin@executivemind.io · executivemind.io

Provider pricing pages retrieved 2026-11-14: platform.claude.com, developers.openai.com, aistudio.google.com, deepseek.com, open.bigmodel.cn, ollama.com/pricing, platform.minimax.io, docs.x.ai, alibabacloud.com, aws.amazon.com/bedrock/pricing, thinkingmachines.ai, console.mistral.ai.

Secondary cross-checks: openrouter.ai, benchlm.ai, cloudzero.com, maximem.ai, requesty.ai, artificialanalysis.ai, vals.ai, tokenpricing.c, truefoundry.com, eesel.ai, venturebeat.com. Where a figure above is marked secondary or flagged, verify against the provider page before external publication or contracting.



EXECUTIVE MIND
EXECUTIVEMIND.IO
ADMIN@EXECUTIVEMIND.IO

LLM Token Pricing Reference Card

November 2026 · Executive Mind, Brisbane · USD per 1 million tokens (input / output) · ordered cheapest input price first

PROVIDER	MODEL	INPUT /1M	OUTPUT /1M	NOTES
Google	Gemini 2.5 Flash-Lite	$0.10	$0.40	Cheapest Gemini
Anthropic	Claude Haiku 5.5	$0.10-varies	$0.50	Two-tier by length
OpenAI	GPT-6 Luna	$0.10	$0.50	Fast $0.20/$1.00
Z.ai	GLM-5.3-Flash	$0.15	$0.50	1M ctx
DeepSeek	deepseek-v4.1-flash	$0.15	$0.60	off-peak; peak in $0.30 / out $1.20
Thinking Machines	Inkling-2-small (Tinker)	$0.30	$1.00	1M ctx; fine-tuning $1.50/M-$3.00/M
MiniMax	MiniMax-M3 (<=512K) (50% off)	$0.30	$1.20	Permanent launch discount
Amazon	Nova 2 Lite (preview)	$0.30	$2.50	Amazon's model; reasoning low/med/high
Z.ai	GLM-5.3-FlashX	$0.37	$1.25	Faster variant
Mistral	Mistral Large 3 (prev)	$0.50	$1.50	Still listed
Google	Gemini 3 Flash (preview)	$0.50	$3.00	1M ctx
DeepSeek	deepseek-v4-pro	$0.66	$1.98	off-peak; peak in $1.32 / out $3.96
Mistral	Mistral Large 4 (preview)	$0.68	$2.09	Preview rate
Moonshot	Kimi K2.7 Code	$0.95	$4.00	256K ctx
OpenAI	GPT-6.1 Terra	$1.00	$5.00	Fast $2/$10
Meta	Muse Spark Open (hosted)	$1.25	$4.25	1M ctx; uniform across Together / OpenRouter / Baseten
Mistral	Mistral Large 4 (list)	$1.36	$4.18	EU/US ×1.1, Priority ×1.75
Z.ai	GLM-5.3 (open weights)	$1.40	$4.40	Open weights; 1M ctx
xAI	Grok 4.7 (flagship)	$2.00	$6.00	500K ctx
Alibaba	Qwen3.8-Max (flagship)	$2.00	$6.00	1M ctx
Anthropic	Claude Sonnet 5.5	$2.00	$10.00	
Google	Gemini 3.1 Pro (promo)	$2.00 / $4.00	$10->$20	Promo; $4.00 from Dec 1
Moonshot	Kimi K3.1 Pro (new Nov 2026)	$2.00	$10.00	1M ctx; Highspeed 2×
OpenAI	GPT-6.1 Sol	$2.00	$10.00	Fast $4/$20
Amazon	Nova Premier	$2.50	$12.50	Amazon's model
Anthropic	Claude Sonnet 4.6 / 4.5	$3.00	$15.00	200K ctx
Moonshot	Kimi K3 (open weights)	$3.00	$15.00	Open weights
Anthropic	Claude Opus 5.5	$4.00	$20.00	
OpenAI	GPT-5.6 Sol (promo)	$4.00	$20.00	Promo to 2026-11-30
Anthropic	Claude Opus 4.5	$5.00	$25.00	200K ctx
Anthropic	Claude Fable 5.2 (flagship)	$8.00	$40.00	Flagship tier
OpenAI	GPT-6 Astra (flagship)	$10.00	$50.00	Batch $5/$25

- **Cheapest real reasoning per token:** GPT-6 Luna ($0.10/$0.50), Gemini 2.5 Flash-Lite ($0.10/$0.40), GLM-5.3-Flash and DeepSeek V4.1-Flash ($0.15/$0.50) — frontier-family reasoning at flash pricing.
- **Batch APIs:** −50% on OpenAI, Anthropic, Google and Bedrock. DeepSeek peak/off-peak is the same lever.
- **Caching:** −90% on Anthropic input; $0.01 on GPT-6 Luna and GLM-5.3-Flash; DeepSeek cache hits around $0.003.
- **Promos expiring 2026-11-30:** Gemini 3.1 Pro doubles to $4.00/$20.00 from Dec 1; GPT-5.6 Sol ends.
- **Ollama Cloud** hosts 20+ open-weight models from $20/mo under zero data retention — a bundling discount on convenience and privacy, not on token price.
- **Verify before contracting.** Published mid-November 2026; provider pricing pages move. This card is a snapshot, not a contract.

For the most recent pricing check **executivemind.io/llm-token-pricing** or email **admin@executivemind.io**.



Compiled by Aria
Web, SEO & Brand
Executive Mind · Brisbane, Queensland

Provider pricing pages retrieved 2026-11-14. Where a figure is marked preview, promo or secondary, verify against the provider page before publishing or contracting. Sources: platform.claude.com, developers.openai.com, aistudio.google.com, deepseek.com, open.bigmodel.cn, ollama.com/pricing, platform.minimax.io, docs.x.ai, alibabacloud.com, aws.amazon.com/bedrock/pricing, thinkingmachines.ai, console.mistral.ai, openrouter.ai, artificialanalysis.ai, vals.ai.