



EXECUTIVE MIND

Built from the beach, for the boardroom.

LLM Token Pricing Comparison — Reasoning-Capable Channels

Reference card · 24 September 2026 · USD per 1M tokens (input / output)

Channel	Model	Input /1M	Output /1M	Notes
OpenAI	GPT-6 Astra (flagship)	\$10.00	\$50.00	Frontier
OpenAI	GPT-6 Sol	\$2.00	\$10.00	Cached in \$0.20
OpenAI	GPT-6 Luna	\$0.10	\$0.50	Cached in \$0.01
Anthropic	Claude Fable 5 (frontier)	\$10.00	\$50.00	
Anthropic	Claude Opus 5	\$5.00	\$25.00	
Anthropic	Claude Sonnet 5	\$2.00	\$10.00	
Anthropic	Claude Haiku 4.5	\$1.00	\$5.00	
Google	Gemini 3.1 Pro	\$2.00	\$12.00	≤200K ctx; \$4/\$18 >200K
Google	Gemini 3.7 Flash	\$0.75	\$3.75	Intro thru Dec 31, then \$1.50/\$7.50
Google	Gemini 2.5 Flash-Lite	\$0.10	\$0.40	Thinking tokens bill as output
DeepSeek	V4-Flash	\$0.14	\$0.28	Cache hit \$0.0028; off-peak ≈ half
DeepSeek	V4-Pro	\$0.435	\$0.87	Cache hit \$0.0036
Z.ai	GLM-5.3-Flash	\$0.15	\$0.50	Cached in \$0.03; 1M ctx; natively multimodal
Z.ai	GLM-5.3	\$1.40	\$4.40	Cached in \$0.26
MiniMax	MiniMax-M3	\$0.30	\$1.20	≤512K ctx; \$0.60/\$2.40 >512K; cache \$0.06
Moonshot AI	Kimi K3	\$3.00	\$15.00	1M ctx; cache-hit in \$0.30; no reasoning surcharge
Moonshot AI	Kimi K2.7 Code	\$0.95	\$4.00	256K ctx; batch ~40%
Moonshot AI	Kimi K2.5	\$0.60	\$3.00	Value tier
OpenRouter	(router — pass-through)	provider rate	+ 5.5% credit fee	BYOK 1M reqs/mo free; some models free
Ollama Cloud	Pro plan	\$20/mo flat	—	Usage-credit pools + GPU-time quotas

Reading the numbers

- Cheapest real reasoning per token: DeepSeek V4-Flash (\$0.14/\$0.28) and GPT-6 Luna (\$0.10/\$0.50) — frontier-family reasoning at flash pricing.
- Batch APIs: ~50% (OpenAI, Anthropic, Google); ~40% (Moonshot, excl. K3). Caching: ~90% input (Anthropic), ~99.9% (DeepSeek hits), \$0.03 (GLM Flash), \$0.30 (Kimi K3).
- Watch: Gemini bills thinking tokens at output rates; DeepSeek official docs now show a peak/off-peak split (peak = 2×).
- OpenRouter = router, not a model: per-token price = underlying provider rate + fee. Value = failover + one API key.
- Prices as published late September 2026 — verify before contracting.