AI API Cost Cheat Sheet

LightCube Data · prices checked 2026-11-24 on each provider's official pricing page · USD per 1M tokens, standard paid tier

How to read it. "1k chat" = 1,000 requests of 2,000 input + 500 output tokens. "1k RAG" = 1,000 requests of 8,000 input + 300 output tokens (long retrieved context, short answer). No caching or batch discount applied. Check the live price before you commit a budget: providers change prices often.

Provider	Model	Input /1M	Cached input /1M	Output /1M	1k chat	1k RAG
OpenAI	gpt-6.2-astra	$8.00	$0.400	$40.00	$36.00	$76.00
OpenAI	gpt-6.2-sol	$2.00	$0.100	$10.00	$9.00	$19.00
OpenAI	gpt-6.1-terra	$1.00	$0.050	$5.00	$4.50	$9.50
OpenAI	gpt-6-luna	$0.10	$0.010	$0.50	$0.45	$0.95
OpenAI	gpt-5.6-luna	$0.20	$0.020	$1.20	$1.00	$1.96
OpenAI	gpt-5.5	$5.00	$0.500	$30.00	$25.00	$49.00
OpenAI	gpt-5.4-mini	$0.75	$0.075	$4.50	$3.75	$7.35
OpenAI	gpt-5.4-nano	$0.20	$0.020	$1.25	$1.03	$1.98
OpenAI	gpt-5-mini	$0.25	$0.025	$2.00	$1.50	$2.60
OpenAI	gpt-5-nano	$0.05	$0.005	$0.40	$0.30	$0.52
OpenAI	gpt-4.1-mini	$0.40	$0.100	$1.60	$1.60	$3.68
OpenAI	gpt-4o-mini	$0.15	$0.075	$0.60	$0.60	$1.38
Anthropic	Claude Fable 5.2	$8.00	$0.200	$40.00	$36.00	$76.00
Anthropic	Claude Opus 5.5	$4.00	$0.200	$20.00	$18.00	$38.00
Anthropic	Claude Opus 4.8	$5.00	$0.500	$25.00	$22.50	$47.50
Anthropic	Claude Sonnet 5.5	$2.00	$0.100	$10.00	$9.00	$19.00
Anthropic	Claude Sonnet 5	$2.00	$0.200	$10.00	$9.00	$19.00
Anthropic	Claude Sonnet 4.6	$3.00	$0.300	$15.00	$13.50	$28.50
Anthropic	Claude Haiku 5.5	$0.10	$0.010	$0.50	$0.45	$0.95
Anthropic	Claude Haiku 4.5	$1.00	$0.100	$5.00	$4.50	$9.50
Google	Gemini 4 Carbon	$2.00	$0.100	$10.00	$9.00	$19.00
Google	Gemini 4 Argon	$2.00	$0.100	$10.00	$9.00	$19.00
Google	Gemini 4 Neon	$3.00	$0.150	$15.00	$13.50	$28.50
Google	Gemini 4 Hydrogen	$1.50	$0.075	$7.50	$6.75	$14.25
Google	Gemini 3.8 Flash	$0.75	$0.075	$3.75	$3.38	$7.12
Google	Gemini 3.5 Flash-Lite	$0.30	$0.030	$2.50	$1.85	$3.15
DeepSeek	deepseek-flash	$0.30	$0.006	$1.20	$1.20	$2.76
DeepSeek	deepseek-v4-pro	$1.32	$0.044	$3.96	$4.62	$11.75

Five ways to cut the bill

1. **Cache the repeated prefix.** System prompts and shared context billed at the cached-input rate cost 90% to 98% less than fresh input on 26 of the 28 models above (see the cached column).
2. **Batch anything that can wait.** Every provider listed with a Batch API bills it at 50% of the standard rate.
3. **Route by difficulty.** Send classification, extraction and short rewrites to the small model; keep the large model for the hard cases.
4. **Cap output tokens.** Output costs 3x to 8x the input price on every row, so a max_tokens limit and "answer in N bullets" instructions save the most.
5. **Trim retrieved context.** In RAG, input dominates: 8 chunks instead of 20 can halve the "1k RAG" column.

Cheapest for the "1k chat" workload

1. OpenAI gpt-5-nano: $0.30
2. OpenAI gpt-6-luna: $0.45
3. Anthropic Claude Haiku 5.5: $0.45
4. OpenAI gpt-4o-mini: $0.60
5. OpenAI gpt-5.6-luna: $1.00

Pricing notes

- **OpenAI gpt-5.5:** Rate for prompts under 272K tokens.
- **Anthropic Claude Haiku 5.5:** Rate for prompts up to 100K tokens; higher above that.
- **Google Gemini 4 Carbon:** Listed price through Nov 30, 2026; $4 / $20 starting Dec 1, 2026.
- **Google Gemini 4 Argon:** Listed price through Nov 30, 2026; $4 / $20 starting Dec 1, 2026.
- **Google Gemini 3.8 Flash:** Listed price through Dec 31, 2026; $1.50 / $7.50 starting Jan 1, 2027. Cache storage billed separately.
- **Google Gemini 3.5 Flash-Lite:** Cache storage billed separately.
- **DeepSeek deepseek-flash:** Peak-hour rate; off-peak is half.
- **DeepSeek deepseek-v4-pro:** Peak-hour rate; off-peak is half.
- OpenAI rows use the short-context rate (prompts up to 272K input tokens).

Sources

- Anthropic: https://platform.claude.com/docs/en/about-claude/pricing
- DeepSeek: https://api-docs.deepseek.com/quick_start/pricing
- Google: https://ai.google.dev/gemini-api/docs/pricing
- OpenAI: https://developers.openai.com/api/docs/pricing

Run your own numbers: the free AI API cost calculator takes your token counts, cache rate and batch share and compares every model above side by side.