

AI API Cost Cheat Sheet

Token Budgeter · prices as of 2026-10-26

Cost per **1,000 typical chats** (600 words in, 150 words out each ≈ 798k input + 200k output tokens). Cheapest 5 per tier.

Budget (output ≤ \$1 per 1M tokens)

Vendor	Model	In \$/1M	Out \$/1M	Per 1k chats
Meta (Llama)	Llama 3.1 8B Instruct	\$0.05	\$0.08	\$0.06
Meta (Llama)	Llama 3.2 1B Instruct	\$0.027	\$0.201	\$0.06
Mistral	Ministral 3 3B 2512	\$0.1	\$0.1	\$0.10
Meta (Llama)	Llama 3.2 3B Instruct	\$0.05	\$0.33	\$0.11
DeepSeek	DeepSeek V4 Flash 0423	\$0.091	\$0.182	\$0.11

Mid-range (output \$1–\$5 per 1M)

Vendor	Model	In \$/1M	Out \$/1M	Per 1k chats
Google	Gemini 3.1 Flash Lite	\$0.25	\$1.5	\$0.50
DeepSeek	DeepSeek V4 Flash 0731	\$0.44	\$1.32	\$0.61
Mistral	Mistral Large 3 2512	\$0.5	\$1.5	\$0.70
Mistral	Devstral 2 2512	\$0.4	\$2	\$0.72
Google	Gemini 3.5 Flash Lite	\$0.3	\$2.5	\$0.74

Premium (output > \$5 per 1M)

Vendor	Model	In \$/1M	Out \$/1M	Per 1k chats
Mistral	Mistral Medium 3.5	\$1.5	\$7.5	\$2.69
xAI	Grok 4.7	\$2	\$6	\$2.79
xAI	Grok 4.6	\$2	\$6	\$2.79
xAI	Grok 4.5	\$2	\$6	\$2.79
Google	Gemini 3.5 Flash	\$1.5	\$9	\$2.99

Cheapest model from each vendor

Vendor	Model	In \$/1M	Out \$/1M	Per 1k chats
Anthropic	Claude Haiku 5.5	\$0.1	\$0.5	\$0.18
DeepSeek	DeepSeek V4 Flash 0423	\$0.091	\$0.182	\$0.11
Google	Gemini 3.1 Flash Lite	\$0.25	\$1.5	\$0.50
Meta (Llama)	Llama 3.1 8B Instruct	\$0.05	\$0.08	\$0.06
Mistral	Ministral 3 3B 2512	\$0.1	\$0.1	\$0.10
OpenAI	GPT-6 Luna Pro	\$0.1	\$0.5	\$0.18
xAI	Grok Build 0.1	\$1	\$2	\$1.20

Prices: list prices from <https://openrouter.ai/api/v1/models>, fetched 2026-10-26T13:14:57+00:00 (55 models tracked; image, moderation and experimental variants omitted). Standard non-batch, non-cached rates; ~1.33 tokens per word. Estimates only: confirm on each provider's pricing page. Live numbers: tokenbudgeter.com