

Model Price & Release Tracker — 2026-10-22

Monthly intelligence for teams running inference in production

Generated 2026-10-22 02:55 UTC from 468 priced models across 232 model families, plus 156 open-weight releases.

7 router/synthetic endpoints with no list price are excluded from the tables below.

Baseline issue. This is the first snapshot, so it establishes the reference point. Price movements appear from the next issue onward, when the tracker has two points to compare.

What changed this month

No prior snapshot to compare against yet.

Newly listed models

Model	Prompt /1M	Completion /1M	Context
recursive/flint-v1.0:free	free	free	262,144
minimax/minimax-m3.1-flash:free	free	free	1,000,000
deepseek/deepseek-v4.1-flash-1017	\$0.3	\$1.2	1,048,576
elorian/glimpse	\$0.6	\$3	262,144
xingchen-agi/xing4.0-29b-a4b:free	free	free	262,144
reflection/beam-1.0-501b-a23b	\$0.3	\$1	1,048,576
inception/mercury-3	\$0.1	\$0.35	1,000,000
moondream/moondream3-preview	\$0.3	\$2.5	32,000
moondream/moondream3.1-9b-a2b	\$0.3	\$2.5	32,000
meta/muse-spark-open	\$1.25	\$4.25	1,048,576
nvidia/nemotron-3.5-super-vl	\$0.15	\$0.5	1,048,576
nvidia/nemotron-3.5-super-vl:free	free	free	1,000,000
unbiased/pareto-26.10	\$0.8	\$3.2	1,048,576
~unbiased/pareto-latest	\$0.8	\$3.2	1,048,576
unbiased/pareto-26.9	\$2.5	\$7.5	262,144

Cheapest models by prompt price

Model	Prompt /1M	Completion /1M	Context
ibm-granite/granite-4.0-h-micro	\$0.017	\$0.112	131,000
openai/gpt-oss-20b	\$0.018	\$0.09	131,072
z-ai/glm-5.3-flash	\$0.02	\$0.121	1,048,576
~z-ai/glm-flash-latest	\$0.02	\$0.121	1,048,576
inclusionai/ling-3.0-flash	\$0.021	\$0.063	262,144
openai/gpt-oss-20b:batch	\$0.024	\$0.112	131,072
nex-agi/nex-n2.5-mini	\$0.025	\$0.1	262,144
openai/gpt-5-nano:batch	\$0.025	\$0.2	400,000
meta-llama/llama-3.2-1b-instruct	\$0.027	\$0.201	60,000
mistralai/mistral-nemo	\$0.029	\$0.03	131,072

Cheapest models for a mixed 3:1 workload

Cost of 1M tokens where 750k are prompt and 250k are completion — the shape of most chat and RAG traffic.

Model	Blended /1M
mistralai/mistral-nemo	\$0.029
scaledown/extract	\$0.03
scaledown/summarize	\$0.03
inclusionai/ling-3.0-flash	\$0.032
openai/gpt-oss-20b	\$0.036
ibm-granite/granite-4.0-h-micro	\$0.041
google/gemma-3-4b-it	\$0.043
sao10k/l3-lunaris-8b	\$0.043
nex-agi/nex-n2.5-mini	\$0.044
z-ai/glm-5.3-flash	\$0.045

Largest context windows

Model	Context	Prompt /1M
x-ai/grok-4.20	2,000,000	\$1.25
x-ai/grok-4.20-multi-agent	2,000,000	\$1.25
meta-llama/llama-4-scout	1,310,720	\$0.1
openai/gpt-5.4	1,050,000	\$2.5
openai/gpt-5.4-pro	1,050,000	\$30
openai/gpt-5.4-pro:batch	1,050,000	\$15
openai/gpt-5.4:batch	1,050,000	\$1.25
openai/gpt-5.5	1,050,000	\$5
openai/gpt-5.5-pro	1,050,000	\$30
openai/gpt-5.5-pro:batch	1,050,000	\$15

New open-weight releases

Model	Released	Downloads	Likes
RecursiveSI/flint	2026-10-21	0	269
InternScience/Sage-A1.5-Q4KM-GGUF	2026-10-21	10	32
ATH-MaaS/OvisOCR3	2026-10-21	7	40
KRAFTON/Raon-1-mini-preview	2026-10-21	25	107
KRAFTON/Raon-1-mini-preview-gguf	2026-10-20	9	26
unsloth/Ornith-1.5-122B-GGUF	2026-10-20	7,758	64
unsloth/flint-GGUF	2026-10-20	11	73
ornith-ai/Ornith-1.5-122B-GGUF	2026-10-20	7,699	37
AtomicChat/DeepSeek-V4.1-Flash-1017-Uncensored-GGUF	2026-10-19	132	30
baselabs/mammoth-1	2026-10-19	21	25
briaai/Fibo-2.0-edit-turbo	2026-10-19	310	87
gradium/phonon-31m-en	2026-10-19	3,590	68
gradium/phonon-40m-multilingual	2026-10-19	349	34
gradium/phonon-40m-en	2026-10-19	562	54
XHToken/Spark-X2.6-4B-GGUF	2026-10-19	1,359	60

[How to read this](#)

Prompt and completion prices are per million tokens, taken from the OpenRouter model catalogue. Open-weight entries are Hugging Face releases with real download traction, so the list is filtered to things people actually use rather than every upload.

A price cut on a model you already depend on is the highest-signal event here: it is a direct margin change, and the window to renegotiate or re-route traffic is usually weeks, not months. A new open-weight release matters if it is cheap enough to self-host at your volume — at that point the per-token price is a GPU-hour question instead of an API question.

Prices captured 2026-10-22. Verify against vendor pricing before committing.