MODEL-AGNOSTIC BY POLICY

The model board

Every major provider's latest and previous model, side by side — price, context, capability, caching semantics, and the operational behaviour that bites at implementation time. One page, so you can stop opening twelve pricing tabs. Every number carries the URL it came from and the date we fetched it.

- Last updated **2026-09-03**
- Refreshed **weekly**
- Next update **2026-09-10**
- **54** models · **12** labs · **3** hosts
- **99** source links, all fetched 2026-09-03
- USD per 1M tokens, list rate — no committed-use, enterprise or volume discount

naderu.com/models-board · Naderu, a venture of BytesBrains Pte. Ltd.

START HERE

Pick by the job, not by the leaderboard.

The right model is rarely the top of any chart — it's the cheapest one that clears your bar. Each card names what we'd reach for and the catch that comes with it. The slug under each card is a published name: this file is served at `/data/model-board.json`, so anything you build can key off the same vocabulary we do.

AGENTIC CODING

Long tool-use loops over whole repos

→ Claude Sonnet 5 Anthropic

→ GPT-5.6 Terra OpenAI

→ Kimi K2.7 Code Moonshot · Kimi

The catch: output tokens dominate an agent loop's bill, and the top of that range is now on a clock: GPT-5.6 Sol's $4.00/$20.00 is labelled promotional through at least 2026-11-21, so the row that currently bills below Claude Opus 5's $25.00 output is the one with an expiry on it. The multiplier itself still varies more than people expect — output runs 5x to 8x the input rate at OpenAI, Anthropic and Google, but 2x to 3x at DeepSeek and xAI. Cache the stable prefix and cap max_tokens before you optimise anything else.

CHAT ASSISTANT

Multi-turn conversation, the ordinary case

→ Claude Haiku 4.5 Anthropic

→ Gemini 3.8 Flash Google · Gemini

→ GPT-5.6 Luna OpenAI

The catch: the trap here is caching rather than headline price: all three cache, but Claude Haiku 4.5 needs 4,096 tokens before a prompt is cacheable at all — four times Claude Sonnet 5's 1,024 — so a short chat system prompt silently pays full rate on every turn, with no error to tell you. This card now points at Gemini 3.8 Flash rather than 3.7: it arrived this pass at exactly the same $0.75/$3.75 and inherits exactly the same clock, doubling to $1.50/$7.50 on 2027-01-01. GPT-5.6 Luna's $0.20/$1.20 carries no end date. Claude Haiku 4.5 does, of a different kind — retirement no sooner than 2026-10-15, the nearest on this board.

EXTRACTION

Classify or pull fields out of a corpus

- → GLM-4.7-FlashX `Z.ai · GLM`
- → DeepSeek-V4-Flash `DeepSeek`
- → Mistral Small 4 `Mistral`

The catch: volume is a throughput constraint, so the job is the same for one row or ten million. Two clocks run under this card. DeepSeek-V4-Flash is shown at its standard peak rate ($0.44/$1.32); off-peak still halves the bill outside 01:00–04:00 and 06:00–10:00 UTC, Monday to Friday. And GLM-5.3-Flash's launch promotion — $0.075/$0.25 against a $0.15/$0.50 list — ends at midnight on 2026-09-09, Singapore time, which is six days after this refresh and before the next one. GLM-4.7-FlashX at $0.07/$0.40 and Mistral Small 4 at $0.15/$0.60 are the only picks here with neither a clock nor a promotion attached, and Mistral Small 4 now carries a sourced cached rate of $0.015 for the first time.

DOCUMENT UNDERSTANDING

OCR, scans, charts and forms

- → Gemini 3.1 Flash-Lite `Google · Gemini`
- → Qwen3.8-Max `Alibaba · Qwen`
- → GPT-5.6 Luna `OpenAI`

The catch: image and audio tokens are counted differently by every provider — Gemini prices audio input at 2x text — so a cost model built on one vendor does not transfer. Measure on your own page sizes. Qwen3.8-Max's vision and video input remain API-only: the open weights released alongside it are text-only, re-confirmed this pass, and the Hugging Face card tags them 'qwen3.8-max' rather than Apache-2.0, so check the licence as well as the modality before planning a self-hosted pipeline.

SUMMARISATION

Long documents in, short faithful text out

- → Gemini 3.8 Flash `Google · Gemini`
- → DeepSeek-V4-Flash `DeepSeek`
- → GLM-4.7-FlashX `Z.ai · GLM`

The catch: this job inverts the agentic-coding arithmetic, and budgets built for that job mislead here. You send a great deal and get back a little, so the input rate is the bill and the output multiplier barely registers — Gemini 3.8 Flash's $0.75 in matters, its $3.75 out mostly does not. Two consequences. First, the cached-input rate is the real lever when you summarise the same corpus repeatedly, and it varies far more than headline input does: 10% of base at Gemini and GLM, but 3.2% at DeepSeek ($0.014 against $0.44). Second, the context window is a hard gate rather than a nice-to-have, and this is exactly where the board's gaps bite — GLM-4.7-FlashX has the cheapest input on the entire board at $0.07, and Z.ai publishes no context window for it at all, which is the one number this job needs before price is even worth discussing. DeepSeek-V4-Flash publishes both (1M in, 384K out) and halves outside 01:00–04:00 and 06:00–10:00 UTC on weekdays, which suits a batch that can wait.

TRANSLATION

Language pairs, with terminology you control

- → Mistral Small 4 `Mistral`
- → Gemini 3.8 Flash `Google · Gemini`
- → Qwen3.7-Plus `Alibaba · Qwen`

The catch: output length tracks input length here — unlike summarisation, you get back roughly what you sent — so the input and output rates both land on the full document and the 5x-to-8x output multiplier applies to all of it. The subtler cost is that a price per token is not a price per word, and the gap widens the further you get from English: the same passage costs materially more tokens in a non-Latin script than its English source, so a per-word budget derived from an English test set will understate the bill on the languages you are actually translating into. Tokenizers also move under you — Anthropic's own pricing page states that Claude 4.7 and later produce roughly 30% more tokens for the same text than earlier models, which is a 30% price rise that never appears in any price column. Measure on your real language pairs, not on English. Mistral Small 4 and Qwen3.7-Plus are both open-weight, so the terminology-control problem can also be solved by fine-tuning rather than by prompting.

COPY EDITING

Rewrite for grammar, tone and house style

- → Claude Sonnet 5 `Anthropic`
- → GPT-5.6 Luna `OpenAI`
- → Mistral Small 4 `Mistral`

The catch: you return the whole document, so this is the text job where the output rate hurts most — the full length is billed at the output multiplier, every time, and a second editing pass costs the same as the first. The style guide is the thing to cache: it is long, it is identical on every call, and it is the whole reason the job works. Watch the minimum cacheable prefix rather than assuming caching applies — Claude Sonnet 5 needs 1,024 tokens before anything caches at all, and a house-style prompt shorter than that pays full rate on every document with no error to tell you. Claude Sonnet 5's $2/$10 is no longer on a clock: the rise to $3/$15 once scheduled for 2026-09-01 was withdrawn, and that date has now passed without it. GPT-5.6 Luna at $0.20/$1.20 is the cheapest serious option here and carries no end date of its own — though everything above it in the GPT-5.6 range now does.

THEN FILTER

A constraint is not a job.

These are the things a request or a deployment has to satisfy — they narrow the table rather than choose from it. Keeping them off the job list is what stops that list sprawling: the entries it would otherwise be missing are products of the two axes, and there is no end to those.

LONG CONTEXT

Windows past a few hundred thousand tokens

the advertised window is not the reliable window, and on Gemini and Grok it is not even the advertised price — both double per token above 200K input, and xAI states the higher rate applies to every token in the request, not just the ones past the line. Anthropic bills its full 1M at standard rates, which flips the ranking on long prompts. Llama 4 Scout's 10M figure stays on shaky ground: the secondary source cited for it has now failed to state that number on three consecutive re-checks, most recently today, and Together AI still lists Scout only for LoRA fine-tuning at $3.00 per 1M tokens rather than for inference.

OPEN WEIGHTS

Weights you can download, run and fine-tune

"open weights" is not one licence, and this pass produced the clearest example yet. GLM-5.3 arrived with downloadable weights tagged "glm-5.3" on Hugging Face — a bespoke licence, where GLM-5.2 immediately below it on this board carries plain MIT. The card does not print the terms, so what it permits is not something we can state; what we can state is that reading "GLM" and assuming MIT is now wrong. Around it: MIT (DeepSeek, GLM-5.2) and Apache-2.0 (the open Qwen lines, and Meta's Muse Glimmer 30B) carry no commercial restriction; Llama and Kimi attach user-count and revenue thresholds; Qwen3.8-Max and GLM-5.3 each ship under a custom licence of their own name; GLM-5.3-Flash states no weights position at all and is carried as unverified rather than assumed; and Codestral sits in Mistral's proprietary Premier tier despite the open-weight history of the name. This is the column Naderu works in — read the licence file, not the badge.

DATA RESIDENCY

Where the tokens are processed, and who can audit it

a region flag is not a compliance story. Mistral is EU-headquartered, Anthropic offers US-pinned inference at a 1.1x premium on every token category, and any MIT-weighted model can run inside your own network — usually the only answer that survives a real review.

LOW LATENCY

Interactive, user-facing turns

time-to-first-token is set by prompt length and cache-hit rate as much as by the model, and any reasoning mode undoes it. Groq remains the only host on this board publishing per-model speeds — 1,000 tok/s on GPT OSS 20B down to 280 on Llama 3.3 70B Versatile — but both its Llama rows are quoted "ContactSales" rather than at a rate, so the fastest published figures sit on the rows with no published price.

HIGH THROUGHPUT

Tokens per second at volume, not per request

throughput is a property of the host, not the model, so it lives in the hosts table rather than the model rows. Groq is the only source on this board that publishes a per-model speed figure at all; every other host is compared on price alone here.

VISION REQUIRED

Image input is not optional

filter on the vision column, then check the same claim twice: vision on the API and vision in the open weights are different statements. Qwen3.8-Max serves vision and video through Model Studio while the checkpoint published on Hugging Face is text-only.

AUDIO REQUIRED

Speech or audio input

few rows state audio input at all, and where it is stated it is priced separately from text: Gemini 3.1 Flash-Lite bills audio input at $0.50 against $0.25 for text, and Gemini 2.5 Flash at $1.00 against $0.30. Budget audio as its own line, not as a modality discount.

TOOLS REQUIRED

Function calling as a hard requirement

this rarely eliminates a model any more — it changes the bill. Tool definitions are input tokens on every request, and Anthropic is the one provider on this board that publishes the overhead: 286 to 474 tokens for the tool-use system prompt depending on model and tool_choice, and about 4,500 more to declare the computer-use toolset.

JSON SCHEMA

Guaranteed structured output

named here so the vocabulary is complete, and so its absence is visible: this board does not yet carry a column for structured-output support. A column ships when every row has an answer sourced from its own provider's page, not before.

NO TRAINING ON YOUR DATA

Contractual, not just a setting

read the retention terms alongside the toggle. Meta ships each Muse Spark release as two separate endpoints — muse-spark-1.3-contributor and muse-spark-1.3, the contributor tier costing an eighth as much precisely because your traffic is used for product improvement — so the guarantee is a model id, not a checkbox, and the cheap row is cheap for a reason worth reading. At the other end, the Claude Fable and Mythos lines carry 30-day retention and are explicitly not available under zero data retention.

BATCH API

Async work at a discount

where it exists it is a flat 50% on both input and output — OpenAI, Anthropic, Google and Mistral all state that rate. Every other lab on this board carries no batch discount in its row, because none of their sourced pages state one. Two exclusions worth knowing before you plan around it: Anthropic's batch discount does not stack with fast mode, and it does not apply to Managed Agents sessions at all.

AND DECIDE

How much you are willing to pay for the last increment.

The same job and the same constraints still leave a choice, and it is an economic one rather than a technical one.

SAVER

Cheapest thing that clears the bar

the large-window end of this posture is where the arithmetic gets slippery. DeepSeek-V4-Flash reads $0.44/$1.32 on this board — its standard peak rate — and halves off-peak, so what you pay depends on when your job runs. GLM-5.3-Flash is cheaper still until midnight on 2026-09-09, Singapore time, and then doubles to its $0.15/$0.50 list without anything being announced; that date falls before the next refresh. Grok 4.3 at $1.25/$2.50 below 200K is unchanged and doubles above it. Kimi K3's flat $3.00/$15.00 across a 1M window is unchanged and first-party confirmed. None of these does vision.

BALANCED

The cheapest row that clears your bar

the default posture, and the one the job cards assume. It is also where most movement on this board lands, and this pass it landed as a free upgrade rather than a cut: Gemini 3.8 Flash arrived at Gemini 3.7 Flash's exact $0.75/$3.75, so the balanced pick got newer without getting dearer. The catch is that it also inherited 3.7's expiry — both double on 2027-01-01 — and GPT-5.6 Sol's $4.00/$20.00, the cut that reordered this tier in August, is now labelled promotional through at least 2026-11-21. Two of the three obvious balanced picks are on a clock.

QUALITY

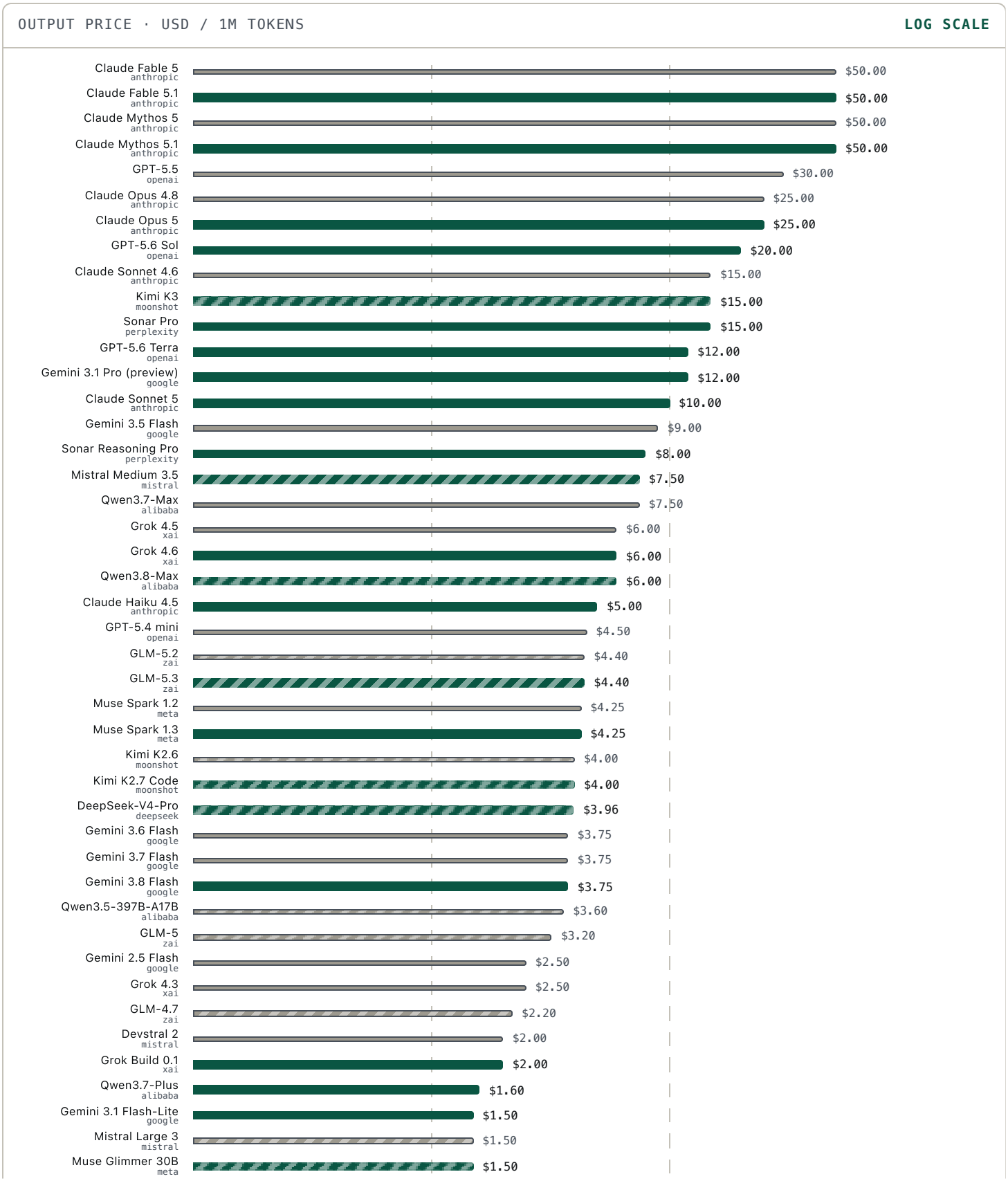
Top of the capability range, bill accepted

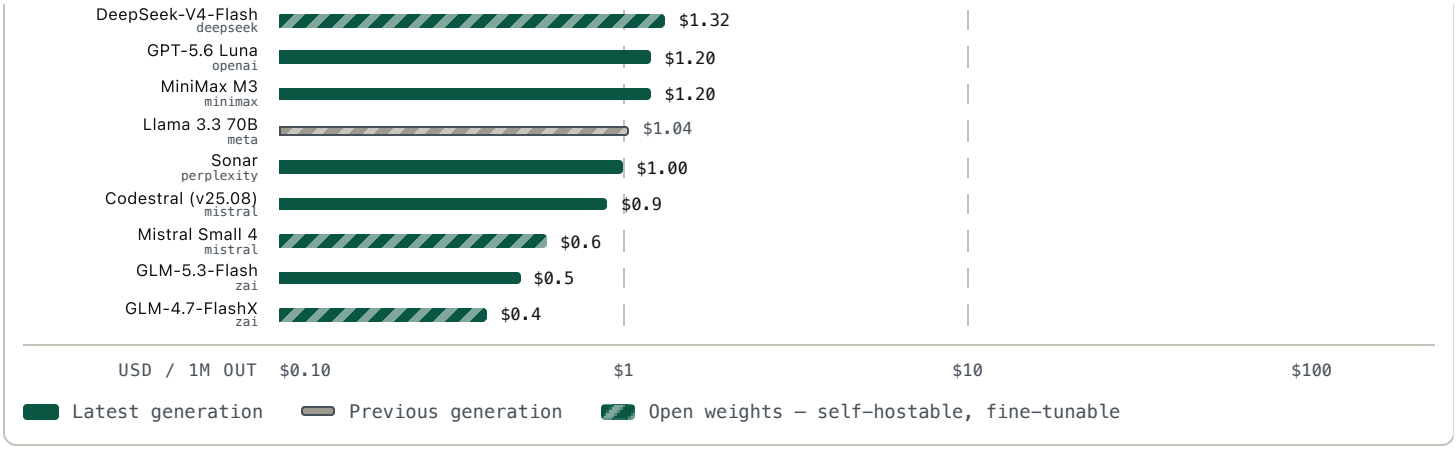
Claude Fable 5.1 and Mythos 5.1 at $10.00/$50.00, Claude Opus 5 at $5.00/$25.00, GPT-5.6 Sol now at $4.00/$20.00. The 5.1 revisions arrived this pass at their predecessors' headline price but with cache reads at 0.025x base instead of 0.1x — $0.25 against $1.00 — so at the top of the board the cheapest move is the newer row, not the older one. Three things to settle before committing: Mythos 5.1 is allocated rather than sold, with no self-serve sign-up, so a roadmap cannot depend on it; the Fable line ships safety classifiers that can decline a request, returning HTTP 200 with stop_reason "refusal" rather than an error, which is an integration change and not a footnote; and GPT-5.6 Sol's rate is promotional through at least 2026-11-21.

COST AT A GLANCE

Output price spans more than 100× across the board.

Output tokens, USD per million, list rate. The axis is logarithmic — the cheap end and the frontier end are too far apart to share a linear scale honestly. Hover any bar for its input and cached-input rates.

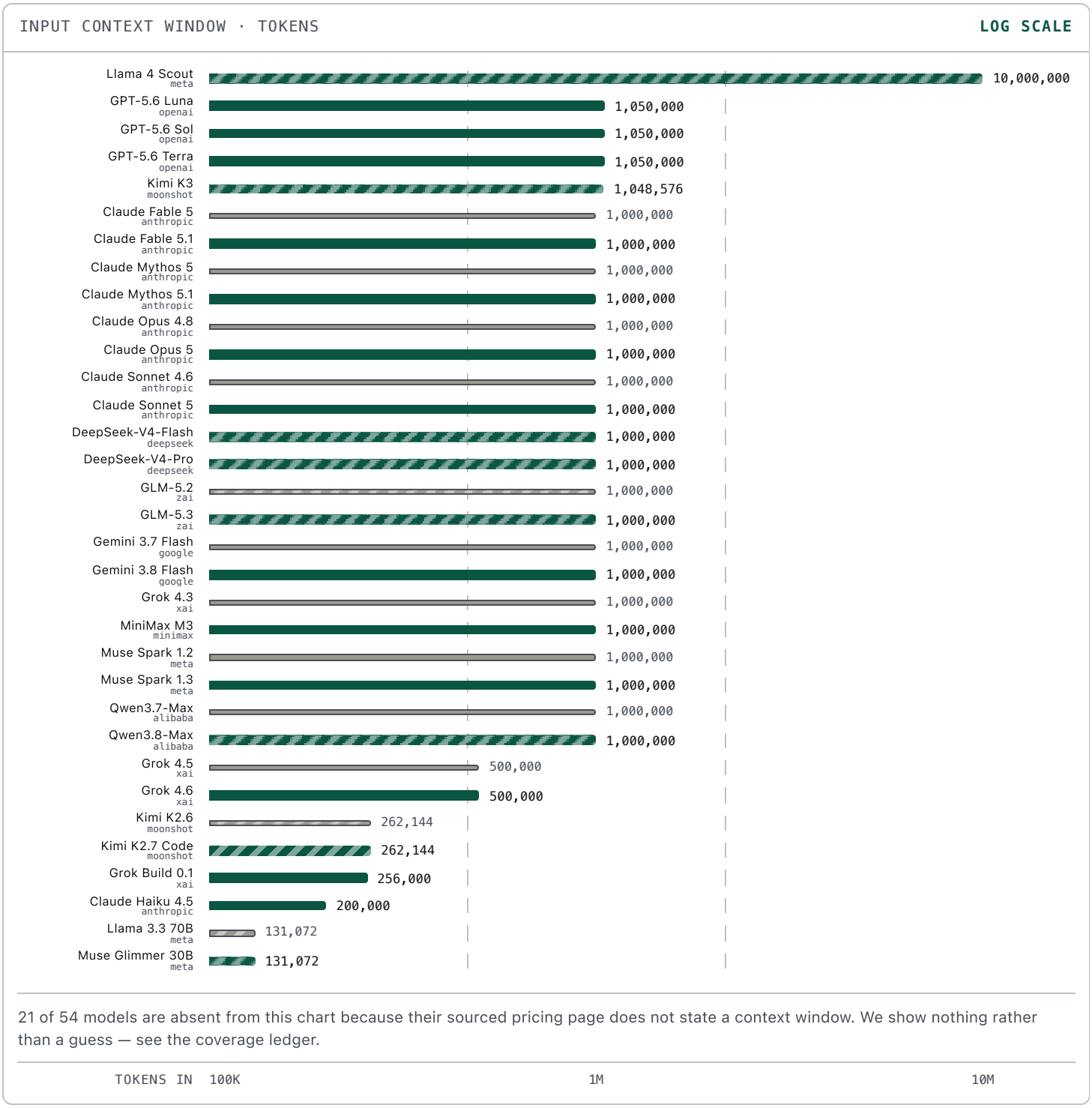



DeepSeek-V4-Flash
deepseek
$1.32
GPT-5.6 Luna
openai
$1.20
MiniMax M3
minimax
$1.20
Llama 3.3 70B
meta
$1.04
Sonar
perplexity
$1.00
Codestral (v25.08)
mistral
$0.9
Mistral Small 4
mistral
$0.6
GLM-5.3-Flash
zai
$0.5
GLM-4.7-FlashX
zai
$0.4
USD / 1M OUT
$0.10
$1
$10
$100
Latest generation
Previous generation
Open weights — self-hostable, fine-tunable

CONTEXT AT A GLANCE

Advertised window ≠ reliable window ≠ affordable window.

Maximum input context, log scale. Three separate caveats apply: recall degrades well before the limit on every model here; Gemini and Grok both double their per-token price above 200K input, so the big window costs twice as much to actually fill; and Anthropic bills its full 1M at standard rates, which reverses the ranking on long prompts.



21 of 54 models are absent from this chart because their sourced pricing page does not state a context window. We show nothing rather than a guess — see the coverage ledger.

THE BOARD

Everything, in one table.

Printed in landscape, all 18 columns, in whatever filter and sort order the page was in. The whole table is in the page's HTML, so the controls only filter and sort and it still reads with JavaScript off. The same data is served as JSON at `/data/model-board.json` if you'd rather query it. A `?` means the sourced page did not state it.

54 of 54 models

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Qwen3.5-397B-A17B ALIBABA · QWEN · PREVIOUS	$0.6	–	**$3.60**	–	–	–	not stated	?	?	Yes	explicit	Open	Apache-2.0 (reported, open Qwen lines)	Large open-weight MoE with the least ambiguous licence here	Apache-2.0 is the least ambiguous licence here, but no cached-input rate is published — cost the full prefix	Model Studio Singapore rate; Alibaba's own page states no cached-input rate. Until 2026-07-30 this row carried $0.35 cached taken from Together, a host; it was withdrawn when the row moved to first-party pricing, because a host's cache rate is not the lab's · not stated on the sourced page	src1 src2 2026-09-03
Qwen3.7-Max ALIBABA · QWEN · PREVIOUS	$2.50	–	**$7.50**	–	1M	65K	effort knob	Yes	Yes	Yes	explicit	Closed	Max tier is closed; smaller Qwen lines are Apache-2.0	Multilingual and multimodal work with a 1M window, if you don't need Qwen3.8-Max's extra scale	Superseded by Qwen3.8-Max as Alibaba's flagship; the headline rate is still a limited-time promotion and both hosts are cheaper — shop it	Model Studio Singapore rate, marked limited-time 50% off list. Alibaba states no per-model cached rate, only multipliers — 10% of input on an explicit cache hit, 20% on an implicit one — so this row still carries no cached figure rather than pick one of the two. Hosts do state one: $1.475/$4.425 on OpenRouter, $1.25/$0.13 cached/$3.75 on Together, and that $0.13 is Together's rate for Together's cache, not Alibaba's	src1 src2 src3 src4 2026-09-03
Qwen3.7-Plus ALIBABA · QWEN · LATEST	$0.4	–	**$1.60**	–	–	–	not stated	?	?	Yes	unverified	Not verified		Cheap general-purpose tier in the Qwen line	Open-weight status not verified this pass — do not assume you can self-host	Model Studio Singapore rate for 0–256K input, marked limited-time 20% off; $1.20 in above 256K. Was carried at $0.32/$1.28 from Together, which no longer lists this model · window not stated; input is priced in two tiers with the step at 256K	src1 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Qwen3.8-Max ALIBABA · QWEN · LATEST	$2.00	$0.25	**$6.00**	—	1M	131K	effort knob	Yes	?	Yes	explicit	Open	custom "qwen3.8-max" licence named on the Hugging Face model card — not Apache-2.0; the card names the licence but does not print its terms, re-checked 2026-09-03	Alibaba's new flagship — a 2.4T-parameter MoE (95B active), now with open text-only weights	Ends a 3-pass deferral on this board, but the API's vision and video input are not in the Hugging Face release — the open checkpoint is text-only despite the hosted API being multimodal. Read the named custom licence before assuming Apache-2.0-style terms	Model Studio Singapore/international rate. The page also shows a region-varying range ($1.65–$2.00 in / $0.206–$0.25 cached / $4.951–$6.00 out); this row uses the international figures for consistency with the rest of the board. No promo/expiry stated on the pricing page, unlike Qwen3.7-Max's · the API states max input 991,808 + max output 131,072; the released weights' native training context is 262,144, extensible to about 1,010,000 per the Hugging Face card — the API figure and the checkpoint's native figure are not the same number	src1 src2 2026-09-03
Claude Fable 5 ANTHROPIC · PREVIOUS	$10.00	$1.00	**$50.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		The top of the board when capability beats cost	Superseded by the 5.1 revision this pass and moved to Legacy on Anthropic's own overview page. Same $10/$50 as 5.1, but cache reads cost 0.1x base here against 5.1's 0.025x — on a cache-heavy workload the newer row is strictly cheaper	1h cache write $20.00 · full 1M window at standard rates — no long-context premium; up to 128K output per request	src1 src2 src3 2026-09-03
Claude Fable 5.1 ANTHROPIC · LATEST	$10.00	$0.25	**$50.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Demanding reasoning and long-horizon agentic work, per Anthropic's own guidance	$50 output is the top of this board. Adaptive thinking is always on, so there is no non-thinking mode to fall back to; effort is tunable and defaults to high. Retirement committed no sooner than 2027-09-01	1h cache write $20.00. The cache read is the number to look at: $0.25 is 0.025x the base input rate, where every other Claude row pays 0.1x. On this row caching is four times the saving it is elsewhere on the board · full 1M window at standard rates; up to 128K output per request	src1 src2 src3 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Claude Haiku 4.5 ANTHROPIC · LATEST	$1.00	$0.1	**$5.00**	−50%	200K	64K	effort knob	Yes	—	Yes	both	Closed		Fast interactive turns that still need real tool use	Retirement is committed no sooner than 2026-10-15 — the nearest date on this board, six weeks out. It also uses the older tokenizer, so token counts are not comparable with 4.7+ models	1h cache write $2.00 · 200K window and 64K max output, both stated on the models overview page this pass — it predates the 4.6+ 1M window guarantee	src1 src2 2026-09-03
Claude Mythos 5 ANTHROPIC · PREVIOUS	$10.00	$1.00	**$50.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Approved Project Glasswing security work — Fable 5's capability with the safety classifiers removed	Superseded by the 5.1 revision this pass and moved to Legacy on Anthropic's own overview page. Same $10/$50 as 5.1, but cache reads cost 0.1x base here against 5.1's 0.025x — on a cache-heavy workload the newer row is strictly cheaper	limited availability; 1h cache write $20.00 · the launch doc states Fable 5 and Mythos 5 share the same specs — 1M window by default, up to 128K output per request	src1 src2 src3 2026-09-03
Claude Mythos 5.1 ANTHROPIC · LATEST	$10.00	$0.25	**$50.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Demanding reasoning and long-horizon agentic work, per Anthropic's own guidance	Limited availability — access is gated, so this rate is not a rate you can simply start paying. $50 output is the top of this board. Adaptive thinking is always on, so there is no non-thinking mode to fall back to; effort is tunable and defaults to high. Retirement committed no sooner than 2027-09-01	1h cache write $20.00. The cache read is the number to look at: $0.25 is 0.025x the base input rate, where every other Claude row pays 0.1x. On this row caching is four times the saving it is elsewhere on the board · full 1M window at standard rates; up to 128K output per request	src1 src2 src3 2026-09-03
Claude Opus 4.8 ANTHROPIC · PREVIOUS	$5.00	$0.5	**$25.00**	−50%	1M	–	effort knob	Yes	—	Yes	both	Closed		Identical pricing to Opus 5 — a safe pin if you are already validated on it	No price advantage over Opus 5, so migrate on capability rather than cost	fast mode $10/$50 · full 1M window at standard rates	src1 src2 2026-09-03
Claude Opus 5 ANTHROPIC · LATEST	$5.00	$0.5	**$25.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Long agent runs and hard multi-step reasoning	New tokenizer on 4.7+ yields ~30% more tokens for the same text — re-measure, don't scale the rate	1h cache write $10.00; fast mode $10/$50 · full 1M window at standard rates; 128K max output per request	src1 src2 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Claude Sonnet 4.6 ANTHROPIC · PREVIOUS	$3.00	$0.3	**$15.00**	−50%	1M	—	effort knob	Yes	—	Yes	both	Closed		The last model on the older tokenizer — useful if your token budgets are tuned to it	Costs 50% more than Sonnet 5 for a generation less capability — and Sonnet 5's rate is now permanent, not introductory	full 1M window at standard rates	src1 src2 2026-09-03
Claude Sonnet 5 ANTHROPIC · LATEST	$2.00	$0.2	**$10.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Best capability-per-dollar on the board, and no longer on a clock	the 50% rise scheduled for 2026-09-01 was withdrawn, and 2026-09-01 has now come and gone at $2/$10 — a budget sized on $3/$15 is overstated	$2/$10 is now the standard rate. The rise to $3/$15 announced for 2026-09-01 will not occur — the pricing page withdrew it on 2026-08-27, and that date has now passed with the rate unmoved. 1h cache write $4.00 · full 1M window at standard rates; 128K max output per request	src1 src2 2026-09-03
DeepSeek-V4-Flash DEEPSEEK · LATEST	$0.44	$0.014	**$1.32**	—	1M	384K	effort knob	—	—	Yes	automatic	Open	MIT (reported)	Bulk extraction, self-hosting, fine-tuning — still a 1M window at a fraction of frontier price	Priced here at the peak/standard rate. Off-peak halves it, so a job you can schedule outside 01:00–04:00 and 06:00–10:00 UTC bills at half this row — and one you cannot schedule bills at this row, not the number quoted in most comparisons	list rate is the peak rate: DeepSeek's own pages call peak standard and off-peak a 50% discount, so this board has shown $0.44/$0.014/$1.32 since 2026-08-27, where it previously showed the off-peak $0.22/$0.007/$0.66. Nothing about DeepSeek's rates moved this pass — only which of the two this board puts in the price column. Off-peak is every hour outside 01:00–04:00 and 06:00–10:00 UTC, Monday to Friday	src1 src2 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
DeepSeek-V4-Pro DEEPSEEK · LATEST	$1.32	$0.044	**$3.96**	—	1M	384K	effort knob	—	—	Yes	automatic	Open	MIT (reported)	Frontier-adjacent capability at a fraction of frontier price, with open weights	Priced here at the peak/standard rate. Off-peak halves it, so a job you can schedule outside 01:00–04:00 and 06:00–10:00 UTC bills at half this row — and one you cannot schedule bills at this row, not the number quoted in most comparisons	list rate is the peak rate: DeepSeek's own pages call peak standard and off-peak a 50% discount, so this board has shown $1.32/$0.044/$3.96 since 2026-08-27, where it previously showed the off-peak $0.66/$0.022/$1.98. Nothing about DeepSeek's rates moved this pass — only which of the two this board puts in the price column. Off-peak is every hour outside 01:00–04:00 and 06:00–10:00 UTC, Monday to Friday	src1 src2 src3 2026-09-03
Gemini 2.5 Flash GOOGLE · GEMINI · PREVIOUS	$0.3	$0.03	**$2.50**	−50%	—	—	effort knob	Yes	Yes	Yes	automatic	Closed		Still one of the cheapest multimodal rows anywhere	Implicit caching is best-effort — you cannot rely on the discount landing	audio input $1.00 · not stated on the pricing page	src1 2026-09-03
Gemini 3.1 Flash-Lite GOOGLE · GEMINI · LATEST	$0.25	$0.025	**$1.50**	−50%	—	—	effort knob	Yes	Yes	Yes	both	Closed		Bulk multimodal extraction at low cost	Audio input is priced at 2x text — separate your modalities when estimating	audio input $0.50 in / $0.05 cached · not stated on the pricing page	src1 2026-09-03
Gemini 3.1 Pro (preview) GOOGLE · GEMINI · LATEST	$2.00	$0.2	**$12.00**	−50%	—	—	effort knob	Yes	Yes	Yes	both	Closed		Documents, video and very large prompts	Doubles per token above 200K input; preview models carry extra rate-limit restrictions	promotional through 2026-12-31, per the pricing page. Above a 200K-token prompt the rate steps to $4.00 in / $0.40 cached / $18.00 out. Cache storage $4.50 per 1M/hr — the most expensive storage rate on this board · not stated on the pricing page; pricing tiers at 200K input	src1 2026-09-03
Gemini 3.5 Flash GOOGLE · GEMINI · PREVIOUS	$1.50	$0.15	**$9.00**	−50%	—	—	effort knob	Yes	Yes	Yes	both	Closed			Now 2x the input and 2.4x the output of 3.6 Flash after that row's cut — strictly worse, by a wider margin than last week	not stated on the pricing page	src1 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Gemini 3.6 Flash GOOGLE · GEMINI · PREVIOUS	$0.75	$0.075	**$3.75**	−50%	—	—	effort knob	Yes	Yes	Yes	both	Closed		The previous Flash workhorse, now priced identically to 3.7 Flash	429 RESOURCE_EXHAUSTED on per-project quota is the failure you will actually hit	introductory on the same timetable as Gemini 3.7 and 3.8 Flash — the pricing page states an identical structure, rising to $1.50/$7.50 on 2027-01-01. Cache storage $0.50 per 1M/hr · not stated on the pricing page	src1 2026-09-03
Gemini 3.7 Flash GOOGLE · GEMINI · PREVIOUS	$0.75	$0.075	**$3.75**	−50%	1M	64K	effort knob	Yes	Yes	Yes	both	Closed		Coding and multi-step agent runs that do not need a frontier price	Superseded by Gemini 3.8 Flash this pass at an identical price, so there is no cost argument for staying. The $0.75/$3.75 is still introductory and still doubles on 2027-01-01	introductory through 2026-12-31; the page states it rises to $1.50/$7.50 on 2027-01-01. Cache storage $0.50 per 1M/hr · 1M window and 64K max output, read from the latest-model page on 2026-08-27. That page now describes Gemini 3.8 Flash instead, so these two figures are carried forward; the prices below were re-read today	src1 src2 2026-09-03
Gemini 3.8 Flash GOOGLE · GEMINI · LATEST	$0.75	$0.075	**$3.75**	−50%	1M	64K	effort knob	Yes	Yes	Yes	both	Closed		The default Flash pick — 3.7's price with a newer model behind it	Arrives at exactly Gemini 3.7 Flash's price, so the upgrade is free, but it inherits the same clock: $0.75/$3.75 is introductory and doubles on 2027-01-01. Thinking is tunable across low, medium and high, and medium is the default — a cost model built on the low setting will not match the bill	introductory through 2026-12-31; the pricing page states it rises to $1.50/$7.50 on 2027-01-01. Cache storage $0.50 per 1M/hr, doubling to $1.00 on the same date. Identical to Gemini 3.7 Flash in every column · 1M window and 64K max output, both stated on the latest-model page	src1 src2 src3 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Llama 3.3 70B META · LLAMA & MUSE · PREVIOUS	$1.04	—	**$1.04**	—	131K	32K	—	—	—	Yes	host-dependent	Open	Llama 3.x Community Licence	Well-understood open-weight baseline with wide host support	Price varies ~2x between hosts for identical weights, and Groq's is now quote-only rather than published — you cannot compare what you cannot see	host price (Together, re-confirmed this pass). Llama 3.3 70B Versatile is back on console.groq.com/docs/models after vanishing last pass, so its availability there is confirmed again — but the row reads "ContactSales" instead of a rate, so Groq still publishes no price for it · max output from Groq's models page (2026-08-06); the window is the model's	src1 src2 2026-09-03
Llama 4 Scout META · LLAMA & MUSE · LATEST	–	–	–	–	10M	–	not stated	Yes	—	Yes	host-dependent	Open	Llama Community Licence — commercial use up to 700M MAU, with name-attribution rules	The largest context window available, self-hostable	No first-party price, limits or caching — everything depends on the host, and Together still lists it only under fine-tuning. The 10M-window claim has now failed re-sourcing three passes running; the licence terms on the same page re-confirm cleanly, so this is the claim that is missing, not the source	no first-party API — price is set by whichever host you pick, and Together AI dropped Llama 4 Scout from its serverless inference table on 2026-08-27; re-checked today, it still appears there only under fine-tuning, at $3.00 per 1M tokens for LoRA SFT with a $6.00 minimum. Fewer hosts list this model than the 10M-window headline suggests · 10M window — the largest on this board; reported by a secondary source last updated May 2026, not confirmed first-party. Re-fetched again this pass: the page still states the licence terms below but still does not state this figure. Carried forward as the best citation available, not freshly verified	src1 2026-09-03
Muse Glimmer 30B META · LLAMA & MUSE · LATEST	$0.35	$0.04	**$1.50**	–	131K	–	effort knob	Yes	?	Yes	host-dependent	Open	Apache-2.0	A genuinely Apache-2.0 Meta model — dense, ~30B, self-hostable	Released 2026-08-10, distilled from Muse Spark by logit distillation. Meta's first Apache-2.0 release since the Llama 4 Community Licence backlash, but there is still no first-party endpoint — this row's price is a host's, not Meta's	host price (Together AI) — Meta does not serve this model first-party; Muse Spark is Meta's only closed model with a first-party endpoint · 131,072+ per the Hugging Face model card, re-read this pass; not stated on Meta's announcement blog	src1 src2 src3 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Muse Spark 1.2 META · LLAMA & MUSE · PREVIOUS	$1.25	$0.15	**$4.25**	–	1M	–	always on	Yes	?	Yes	unverified	Closed		Coding agents on a first-party Meta endpoint, flat-priced across 1M tokens	Superseded by Muse Spark 1.3 this pass at an identical rate; 1.1 and 1.2 are both still on the pricing page. Meta's first closed model — this row is not open weights, and nothing in the Llama licence applies to it	no long-context premium — the same rate across the full window. A `muse-spark-1.2-contributor` tier bills $0.10 in / $0.002 cached / $0.20 out in exchange for your traffic being used for product improvement, at 100 RPM against the standard tier's 3,000. Web-search grounding is $2.50 per 1,000 queries on top · max output not stated on either sourced page	src1 src2 2026-09-03
Muse Spark 1.3 META · LLAMA & MUSE · LATEST	$1.25	$0.15	**$4.25**	–	1M	–	always on	Yes	?	Yes	unverified	Closed		Coding agents on a first-party Meta endpoint, flat-priced across 1M tokens	The contributor tier is an eighth of the price because it is a different trade: your traffic is used for product improvement, at 100 RPM against the standard tier's 3,000. Meta's model pages sit behind a login — the two sources here are the public ones	identical to 1.2 and 1.1, which remain on the pricing page. No long-context premium — the same rate across the full window. A `muse-spark-1.3-contributor` tier bills $0.10 in / $0.002 cached / $0.20 out in exchange for your traffic being used for product improvement · max output not stated on either sourced page	src1 src2 2026-09-03
MiniMax M3 MINIMAX · LATEST	$0.3	$0.06	**$1.20**	–	1M	262K	not stated	?	?	Yes	explicit	Not verified		Cheap general-purpose model with an 80% cached-input discount	The cheap rate is promotional and only holds below 512K input; open-weight status unverified	host price (Together). OpenRouter still lists $0.23/$0.96 with no discount label, and now also states a $0.05 cache-read rate not previously on this board. The 512K input step recorded on 2026-07-30 was not restated on either page again this pass · 1M advertised, but priced in two tiers — the step is at 512K input, not at the window. OpenRouter states 1,048,576 input and 262,144 max completion tokens; the max output is new to this row this pass	src1 src2 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Codestral (v25.08) MISTRAL · LATEST	$0.3	$0.03	**$0.9**	−50%	—	—	—	—	?	?	unverified	Closed	Mistral Premier tier — proprietary, per the models doc	Inline code completion and fill-in-the-middle, where latency matters more than reasoning	Added on 2026-08-27 to close a real gap: this board tracked agentic coding but had nowhere to put a completion model, and they are different purchases — FIM capability, a much tighter latency budget, and no reasoning mode. Despite the open-weight history of the Codestral name, the current v25.08 is listed under Mistral's Premier tier, which is proprietary	cached input $0.03 per 1M, printed per model on docs.mistral.ai/inference/pricing — the third Mistral pricing page, added to this row's sources this pass. It is the −90% discount the other two pages describe only in prose · not stated on either sourced page	src1 src2 src3 2026-09-03
Devstral 2 MISTRAL · DEPRECATED	$0.4	—	**$2.00**	−50%	—	—	not stated	—	—	Yes	unverified	Mixed	listed as "Open" on the pricing page; specific licence not stated	Nothing new — retired, and listed here only so the retirement is on the record	Retired on 2026-06-30. Migrate to Mistral Medium 3.5	retired 2026-07-31, deprecated 2026-05-22, with Mistral Medium 3.5 named as the migration target — both dates read from docs.mistral.ai on 2026-08-27, correcting the 2026-06-30 retirement date this board carried until then. The $0.40/$2.00 shown is the last rate recorded while the model was listed; it is off the pricing page entirely now, so there is no current rate to verify. This board carried it as current for two passes and was wrong to. Re-checked on docs.mistral.ai/inference/pricing this pass, the page that prints cached input for every live Mistral model: Devstral 2 is absent from it, consistent with retirement · not stated on the sourced pricing page	src1 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Mistral Large 3 MISTRAL · PREVIOUS	$0.5	$0.05	**$1.50**	−50%	—	—	not stated	Yes	—	Yes	unverified	Open	Apache-2.0 (v25.12)	Cheaper than Medium 3.5 in both directions — check whether you need the newer model	The previous flagship now undercuts the current one; verify capability on your eval	cached input $0.05 per 1M, printed per model on docs.mistral.ai/inference/pricing — the third Mistral pricing page, added to this row's sources this pass. It is the −90% discount the other two pages describe only in prose · not stated on the sourced pricing page	src1 src2 src3 2026-09-03
Mistral Medium 3.5 MISTRAL · LATEST	$1.50	$0.15	**$7.50**	−50%	—	—	not stated	Yes	—	Yes	unverified	Open	Modified MIT (v26.04)	Agentic and coding work with EU data residency	This board carried "no caching" here until 2026-08-29 on the strength of a rate missing from the pages it cited, retracted it, then left cached_in null for one pass. Both are now closed: the rate was on a third Mistral page all along, and it is $0.15	cached input $0.15 per 1M, printed per model on docs.mistral.ai/inference/pricing — the third Mistral pricing page, added to this row's sources this pass. It is the −90% discount the other two pages describe only in prose · not stated on the sourced pricing page	src1 src2 src3 2026-09-03
Mistral Small 4 MISTRAL · LATEST	$0.15	$0.015	**$0.6**	−50%	—	—	effort knob	Yes	—	Yes	unverified	Open	Apache-2.0 (v26.03)	Our most-reached-for fine-tuning base — cheap, small, open lineage	Hybrid instruct/reasoning/coding model — pin the mode or output length varies	cached input $0.015 per 1M, printed per model on docs.mistral.ai/inference/pricing — the third Mistral pricing page, added to this row's sources this pass. It is the −90% discount the other two pages describe only in prose · not stated on the sourced pricing page	src1 src2 src3 2026-09-03
Kimi K2.6 MOONSHOT · KIMI · PREVIOUS	$0.95	$0.16	**$4.00**	—	262K	—	not stated	?	?	Yes	explicit	Open	Modified MIT (reported)	Cheaper general-purpose alternative to K3, now priced the same as K2.7 Code	No longer more expensive than K2.7 Code — pick on coding-vs-general fit, not price	down from $1.20/$0.20/$4.50 (host-sourced) now that first-party pricing is citable again	src1 2026-09-03
Kimi K2.7 Code MOONSHOT · KIMI · LATEST	$0.95	$0.19	**$4.00**	—	262K	—	not stated	?	?	Yes	explicit	Open	Modified MIT (reported)	Coding agents — a third of K3's price with cached input at 20%	Coding-specialised, so check general reasoning on your own suite. Now first-party sourced, and its context window is documented for the first time: 262,144, well short of K3's 1M	a dearer 'highspeed' variant is also listed at $1.90/$0.38/$8.00 — this row is the standard tier	src1 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Kimi K3 MOONSHOT · KIMI · LATEST	$3.00	$0.3	**$15.00**	—	1.05M	—	not stated	?	?	Yes	explicit	Open	Modified MIT (reported) — thresholds above ~$20M MaaS revenue and 100M MAU	A 1M window with no long-context premium, and weights you can host	Now first-party sourced and unchanged in price from the prior host-sourced figure	flat across the full window — no length tiering	src1 src2 2026-09-03
GPT-5.4 mini OPENAI · PREVIOUS	$0.75	$0.075	**$4.50**	−50%	—	—	not stated	?	?	Yes	automatic	Closed		Cheap tier if you are already pinned to it	GPT-5.6 Luna is dearer per token but a generation newer — compare on your own eval	not stated on the sourced pricing page	src1 2026-09-03
GPT-5.5 OPENAI · PREVIOUS	$5.00	$0.5	**$30.00**	−50%	—	—	not stated	?	?	Yes	automatic	Closed			Now dearer than GPT-5.6 Sol on both sides — $5.00/$30.00 against Sol's $4.00/$20.00 after the 2026-08-22 cut — for a generation less capability. Nothing recommends it for new work	not stated on the sourced pricing page	src1 2026-09-03
GPT-5.6 Luna OPENAI · LATEST	$0.2	$0.02	**$1.20**	−50%	1.05M	128K	effort knob	Yes	Yes	Yes	automatic	Closed		High-volume work that still needs vision and tools	An 80% cut drops a frontier lab into open-model pricing — re-derive any routing rule that sends bulk work elsewhere on price alone	down 80% from the $1.00/$0.10/$6.00 this board recorded on 2026-07-30; the long-context tier is $0.40/$0.04/$0.50/$1.80. The page names a 272K threshold for gpt-5.5/5.4 but not for the 5.6 series	src1 src2 2026-09-03
GPT-5.6 Sol OPENAI · LATEST	$4.00	$0.4	**$20.00**	−50%	1.05M	128K	effort knob	Yes	Yes	Yes	automatic	Closed		The hardest reasoning and longest agent runs	The cheapest it has ever been, and now on a clock: OpenAI calls $4.00/$20.00 promotional through at least 2026-11-21. Budget the rate you would pay if it reverted. Its output now bills below Claude Opus 5's $25.00, having billed above it until August	cut on 2026-08-22 from $5.00/$0.50/$6.25/$30.00. The page now labels $4.00/$20.00 promotional and says it holds "at least through November 21, 2026" — new wording this pass; the rate itself did not move. The long-context tier is $8.00/$0.80/$10.00/$30.00. The page names a 272K threshold for the gpt-5.5 and gpt-5.4 rows but still states none for the 5.6 series, so the point that tier starts is undocumented	src1 src2 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
GPT-5.6 Terra OPENAI · LATEST	$2.00	$0.2	**$12.00**	−50%	1.05M	128K	effort knob	Yes	Yes	Yes	automatic	Closed		The default working model — coding, tools, long context	Reasoning effort silently multiplies the output bill; pin it explicitly	down from the $2.50/$0.25/$15.00 this board recorded on 2026-07-30; the long-context tier is $4.00/$0.40/$5.00/$18.00. The page names a 272K threshold for gpt-5.5/5.4 but not for the 5.6 series	src1 src2 2026-09-03
Sonar PERPLEXITY · SONAR · LATEST	$1.00	–	**$1.00**	–	–	–	—	—	—	—	none	Closed		Search-grounded answers with citations, no retrieval stack to build	The per-request search fee usually exceeds the token cost — model requests, not tokens. Separately: the pricing page now says "Sonar Chat Completions is now Agent API" and that Sonar "will be supported until September 27, 2026". That date is on the Chat Completions endpoint, not on this model — no Sonar model is marked deprecated — but if you call it through Chat Completions you have a migration, not a renewal	plus $5–$12 per 1,000 requests depending on search-context size · not stated on the pricing page	src1 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Sonar Pro PERPLEXITY · SONAR · LATEST	$3.00	—	**$15.00**	—	—	—	—	—	—	—	none	Closed		Deeper grounded research with more sources per answer	$15 output plus request fees — as dear as a frontier model for a narrower job. Separately: the pricing page now says "Sonar Chat Completions is now Agent API" and that Sonar "will be supported until September 27, 2026". That date is on the Chat Completions endpoint, not on this model — no Sonar model is marked deprecated — but if you call it through Chat Completions you have a migration, not a renewal	plus $6–$14 per 1,000 requests · not stated on the pricing page	src1 2026-09-03
Sonar Reasoning Pro PERPLEXITY · SONAR · LATEST	$2.00	—	**$8.00**	—	—	—	always on	—	—	—	none	Closed		A reasoning-capable grounded tier between Sonar and Sonar Pro	Added on 2026-08-27 — not yet compared against Sonar Pro on your own eval; the per-request fee still usually exceeds the token cost. Separately: the pricing page now says "Sonar Chat Completions is now Agent API" and that Sonar "will be supported until September 27, 2026". That date is on the Chat Completions endpoint, not on this model — no Sonar model is marked deprecated — but if you call it through Chat Completions you have a migration, not a renewal	plus $6/$10/$14 per 1,000 requests by low/medium/high search-context size — the same request-fee schedule as Sonar Pro · not stated on the pricing page	src1 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
Grok 4.3 XAI · GROK · PREVIOUS	$1.25	$0.2	**$2.50**	—	1M	—	not stated	?	?	Yes	explicit	Closed		One of the cheapest 1M-window rows on the board	A previous-gen model with a larger window and lower price than the current one — read that carefully	at/above 200K input: $2.50 in / $0.40 cached / $5.00 out · max output not stated on the models page	src1 2026-09-03
Grok 4.5 XAI · GROK · PREVIOUS	$2.00	$0.3	**$6.00**	—	500K	—	not stated	?	?	Yes	explicit	Closed		Mid-price frontier with a 500K window	Everything doubles at 200K input — the big window costs 2x to actually use	at/above 200K input: $4.00 in / $0.60 cached / $12.00 out. Superseded by Grok 4.6 on 2026-08-12 at the same headline rate, but 4.5 keeps the cheaper $0.30 cached-input rate · max output not stated on the models page	src1 2026-09-03
Grok 4.6 XAI · GROK · LATEST	$2.00	$0.5	**$6.00**	—	500K	—	effort knob	Yes	—	Yes	automatic	Closed		Long-running agents and codebase work under a 200K prompt	the cache discount got worse, not better — a prefix-heavy loop may still bill lower on 4.5	at/above 200K input: $4.00 in / $1.00 cached / $12.00 out. Same headline rate as Grok 4.5, but cached input is $0.50 against 4.5's $0.30 — the successor's cache costs 67% more · max output not stated on the models page	src1 2026-09-03
Grok Build 0.1 XAI · GROK · LATEST	$1.00	$0.2	**$2.00**	—	256K	—	not stated	?	?	Yes	explicit	Closed		Coding-oriented tier at the low end of the price range	Smaller 256K window than the general models, same 200K price step	at/above 200K input: $2.00 in / $0.40 cached / $4.00 out	src1 2026-09-03
GLM-4.7 Z.AI · GLM · PREVIOUS	$0.6	$0.11	**$2.20**	—	—	—	effort knob	—	—	Yes	explicit	Open	MIT (reported)	Well-priced open-weight middle ground	Superseded by 4.7-FlashX at a tenth the input price for lighter work	not stated on the pricing page	src1 2026-09-03
GLM-4.7-FlashX Z.AI · GLM · LATEST	$0.07	$0.01	**$0.4**	—	—	—	effort knob	—	—	Yes	explicit	Open	MIT (reported)	The cheapest first-party row on the board — bulk classification and extraction	A free Flash tier sits alongside it; test whether you need to pay at all	GLM-4.7-Flash is free on the API · not stated on the pricing page	src1 2026-09-03
GLM-5 Z.AI · GLM · PREVIOUS	$1.00	$0.2	**$3.20**	—	—	—	effort knob	—	—	Yes	explicit	Open	MIT (reported)	Cheaper than 5.2 in both directions if the capability clears your bar	Verify on your own eval — the price gap to 5.2 is small	not stated on the pricing page	src1 2026-09-03

MODEL	PRICE · USD / 1M				CONTEXT		CAPABILITY					OWNERSHIP		JUDGEMENT			SOURCED
PROVIDER · MODEL	IN	CACHED IN	OUT	BATCH	WINDOW	MAX OUT	REASONING	VISION	AUDIO IN	TOOLS	CACHING	WEIGHTS	LICENCE	BEST FOR	WATCH OUT	NOTES	SOURCE
GLM-5.2 Z.AI · GLM · PREVIOUS	$1.40	$0.26	**$4.40**	–	1M	131K	effort knob	—	—	Yes	explicit	Open	MIT (reported)	MIT-licensed frontier-adjacent coding model you can self-host	Superseded by GLM-5.3 this pass at the same $1.4/$4.4. The reason to care which one you pull is the licence, not the price: GLM-5.2's weights are reported MIT, GLM-5.3's are not	secondary source (2026-07-18); the first-party pricing page states no window	src1 src2 2026-09-03
GLM-5.3 Z.AI · GLM · LATEST	$1.40	$0.26	**$4.40**	–	1M	128K	always on	—	—	Yes	unverified	Open	GLM-5.3 Licence — a bespoke licence, not MIT	The current GLM flagship, with a first-party 1M window on the record	Read the licence before you ship. The Hugging Face card tags this "glm-5.3", a bespoke licence — not the MIT that GLM-5.2 carries — and the card does not print the terms, so what it permits is not something this board can state. Text-only: no vision, unlike much of the rest of this price band	same rate as GLM-5.2, which it supersedes. Cached-input storage is free for a limited time, per the pricing page · 1M window and 128K max output, both stated first-party on docs.z.ai — GLM-5.2's window is only second-hand	src1 src2 src3 2026-09-03
GLM-5.3-Flash Z.AI · GLM · LATEST	$0.15	$0.03	**$0.5**	–	–	–	not stated	?	?	?	explicit	Not verified		The cheapest tier in the GLM line while the launch promotion holds	Added on 2026-08-27 and priced on a promotion with a published end date — size a budget on the $0.15/$0.50 list rate, not the $0.075/$0.25 you will be billed until 2026-09-09. Open-weight status is not stated on the pricing page, so do not assume the rest of the GLM line's MIT terms carry over	list rate. A launch promotion halves it to $0.075/$0.015/$0.25 until "24:00 on September 9, 2026 (UTC+8, Singapore time)" — at the promotional rate this is the cheapest output on the board, and on 2026-09-10 it stops being that without anything being announced · not stated on the pricing page	src1 2026-09-03

All 18 columns are printed. The Source column keeps each row's link labels and fetch date; the URLs behind them are on naderu.com/models-board and in /data/model-board.json.

CACHING & THE REAL BILL

The sticker price is not the invoice.

Most production workloads resend the same long prefix — a system prompt, a schema, a document — on every call. Whether the provider lets you pay once for that is worth more than a few cents on the headline rate. Below: the same job on each provider, **a 50,000-token prompt reused ten times** with 1,000 output tokens each. These figures are computed from the rate table above, not typed in, so they cannot drift away from it.

SAME JOB, TEN CALLS · 50K PROMPT + 1K OUTPUT EACH — COMPUTED FROM THE RATES ABOVE

MODEL	CACHE TYPE	TTL	WRITE RATE	10 CALLS, NO CACHE	10 CALLS, CACHED	YOU SAVE
DeepSeek-V4-Flash CACHE HIT $0.007 VS MISS $0.22 OFF-PEAK (~3%); REASONING EFFORT IS NOW DOCUMENTED IN THREE NAMED LEVELS — LOW, HIGH, MAX — RATHER THAN A FLAT 'CONTROLLABLE', AND THE API ADDED NATIVE OPENAI RESPONSES API SUPPORT WITH ONE-CLICK CODEX INTEGRATION	automatic	—	no premium	$0.233	**$0.042**	82%
GPT-5.6 Terra CACHED INPUT AT 10% OF BASE	automatic	—	$2.50	$1.120	**$0.335**	70%
Gemini 3.7 Flash CACHED INPUT AT 10% OF BASE; EXPLICIT CACHES ALSO BILL HOURLY STORAGE	both	—	no premium	$0.412	**$0.109**	74%
Claude Opus 5 READ 0.1X INPUT; 5M WRITE 1.25X, 1H WRITE 2X	both	5m / 1h	$6.25	$2.750	**$0.787**	71%
Kimi K3 CACHE HIT AT 10% OF INPUT	explicit	—	no premium	$1.650	**$0.435**	74%
Mistral Medium 3.5 A PER-MODEL CACHED-INPUT RATE IS PUBLISHED, BUT NO PAGE THIS ROW CITES SAYS WHETHER CACHING IS AUTOMATIC OR EXPLICIT, STATES A TTL, OR NAMES A MINIMUM PREFIX — SO THE MECHANICS STAY UNVERIFIED EVEN THOUGH THE RATE NO LONGER IS	unverified	—	—	$0.825	**$0.825**	0%

Read the last two columns together. **Mistral Medium 3.5 is cheaper than Claude Opus 5 per token and dearer than it for this job** — but only because Opus 5's cached rate sits on a page this board cites and Mistral's does not, so the Mistral row is priced here with no cache at all. Read that 0% as unpriced, not as absent; the rate lands on 2026-09-03. The inversion this panel exists to show is real — it is just not this row's to demonstrate this week.

BENEFITS & CHALLENGES

What bites you, per provider.

The failures that cost you a weekend are rarely capability failures — they're quota, caching, tokenizer and licensing failures. Every item below is behaviour stated in the provider's own documentation, linked so you can check it. Where we're relying on a secondary report, the link says so.

OpenAI

BENEFIT The widest tooling and SDK surface, a 50% Batch API discount on every listed model, and automatic prefix caching at 10% of the input rate — you get the discount without managing breakpoints. source https://developers.openai.com/api/docs/pricing

BLOCKER Throughput is gated by cumulative paid spend, not by your plan: Tier 1 at $5 paid through Tier 5 at $1,000 paid. Limits apply across RPM, TPM, RPD, TPD and IPM at once and whichever trips first wins — so a launch can meet a ceiling you cannot raise that day. source https://developers.openai.com/api/docs/guides/rate-limits

WATCH Read headroom off the `x-ratelimit-remaining-requests` and `x-ratelimit-reset-requests` response headers rather than inferring it — the tier tables are monthly spend caps, not per-minute capacity. source https://developers.openai.com/api/docs/guides/rate-limits

WATCH Reasoning effort is adjustable from none to max, and reasoning tokens bill as output. A cheap model at max effort can cost more per answer than a dear one at none. source https://developers.openai.com/api/docs/models

---

Mitigation: Qualify your tier weeks before launch, back off exponentially, and meter against the x-ratelimit-* headers.

Anthropic

BENEFIT The full 1M-token window is billed at standard per-token rates on Claude 4.6 and later — a 900K request costs the same per token as a 9K one. Gemini and Grok both double their rate above 200K input, so on long prompts the headline gap narrows or reverses. source https://platform.claude.com/docs/en/about-claude/pricing

BENEFIT Caching is deterministic and cheap to read: a cache hit is 0.1x the input rate. You can let it manage breakpoints automatically with one top-level `cache_control`, or place them by hand. source https://platform.claude.com/docs/en/about-claude/pricing

BLOCKER Claude Fable 5 ships safety classifiers that can decline a request outright. A refusal comes back as a successful HTTP 200 with `stop_reason: "refusal"` and the name of the classifier that fired — not an error — so an integration that only checks the status code reads a refusal as an empty answer. Claude Mythos 5 is the same model with the classifiers removed, and it is not generally available: it is limited to approved Project Glasswing customers. source https://platform.claude.com/docs/en/about-claude/models/introducing-claude-fable-5-and-claude-mythos-5

WATCH Budget for refusal handling on Fable 5, not just for tokens. Anthropic does not bill a request refused before any output, and offers three retry routes — a server-side `fallbacks` parameter (beta), SDK middleware, or your own retry — with fallback credit refunding the prompt-cache cost of switching models so you do not pay it twice. source https://platform.claude.com/docs/en/about-claude/models/introducing-claude-fable-5-and-claude-mythos-5

WATCH Claude 4.7 and later use a newer tokenizer that produces roughly 30% more tokens for the same text. Per-token rates therefore understate the cost change when you migrate from 4.6 or compare against another vendor — re-measure on your own prompts, don't scale the rate. source https://platform.claude.com/docs/en/about-claude/pricing

BENEFIT Sonnet 5's $2/$10 is now the standard rate. The 50% increase to $3/$15 announced for 2026-09-01 has been withdrawn — this board carried it as a scheduled change for two passes, and it is off. Budget today's rate. source https://platform.claude.com/docs/en/about-claude/pricing

WATCH A 5-minute cache write costs 1.25x input and a 1-hour write 2x, so the 5m tier pays off after one read and the 1h tier after two. Bursty low-QPS traffic can miss enough to lose money on the write. source https://platform.claude.com/docs/en/about-claude/pricing

WATCH Pinning inference to the US with `inference_geo: "us"` applies a 1.1x multiplier to every token category including cache reads and writes. source https://platform.claude.com/docs/en/about-claude/pricing

WATCH Fable 5 has already been withdrawn once. Anthropic suspended access for all users on 2026-06-12 under US export controls it could not enforce per-user, and restored it on 2026-07-01 once those lifted, alongside a new classifier for the bypass that triggered them. It is generally available today; the episode is the reason to keep a fallback model configured rather than assume continuity. source https://www.anthropic.com/news/redeploying-fable-5

Mitigation: One breakpoint after the stable prefix, cache-hit rate on a dashboard, a fallback model configured, and re-count tokens after any model migration.

Google · Gemini

BENEFIT The cheapest credible route to a large window, a 50% batch discount on every model, and both implicit and explicit context caching. source https://ai.google.dev/gemini-api/docs/pricing

BLOCKER Rate limits are enforced per project, not per API key, and evaluated against RPM, TPM and RPD simultaneously — exceeding any single one returns `429 RESOURCE_EXHAUSTED`. One backfill job on the same project starves your production traffic, and issuing a second key does not help. source https://ai.google.dev/gemini-api/docs/rate-limits

BLOCKER The actual RPM/TPM numbers are not in the docs — they live in the AI Studio rate-limit dashboard. You cannot capacity-plan from the documentation alone, which is exactly when the 429s arrive. source https://ai.google.dev/gemini-api/docs/rate-limits

WATCH Tiers are spend-gated (Tier 1 needs an active billing account and caps at $250, with a further $10-per-10-minutes cap; Tier 3 needs $1,000 cumulative spend and 30 days). source https://ai.google.dev/gemini-api/docs/rate-limits

WATCH Pricing steps up above 200K input tokens: Gemini 3.1 Pro goes from $2/$12 to $4/$18 per 1M. A long prompt costs double per token, not merely more tokens. source https://ai.google.dev/gemini-api/docs/pricing

WATCH Explicit context caches bill storage by the hour ($1.00–$4.50 per 1M tokens per hour) on top of the read rate. A cache you forgot to delete is a standing line item. source https://ai.google.dev/gemini-api/docs/pricing

---

Mitigation: Separate GCP projects per environment, a client-side token-bucket limiter, backoff with jitter, and a second provider behind a feature flag.

Mistral

BENEFIT EU-headquartered with European data-residency options, and the cheapest credible small models on the board — Mistral Small 4 at $0.15/$0.60 per 1M. source https://mistral.ai/pricing/api/

BENEFIT Several genuinely open-weight lines, which makes them real fine-tuning bases rather than just endpoints. source https://docs.mistral.ai/getting-started/models/

WATCH The API pricing page states a –90% cached-input discount across the API but breaks it out per model only for the resold GLM 5.2 row. This board read that silence as "no prompt caching" and billed the consequence in full — a claim no sourced page makes. Retracted 2026-08-29; see the changelog for what replaces it. source https://mistral.ai/pricing/api/

WATCH Licences vary per model across Apache-2.0, Modified MIT, CC BY-NC 4.0 and a commercial 'Premier' tier. Read the individual model card before building on any of the weights — the family name tells you nothing. source https://docs.mistral.ai/getting-started/models/

WATCH Mistral now resells Z.ai's GLM 5.2 on its own API at $1.4 in / $0.14 cached / $4.4 out — a cached rate below Z.ai's own $0.26, on a model Mistral did not train. A European endpoint for Chinese open weights is a real residency option, but the row you are buying is a third party's model on Mistral's bill. source https://mistral.ai/pricing/api/

Mitigation: Until a per-model cached rate is sourced onto this board, price a prefix-heavy Mistral workload from Mistral's own docs, not from these rows.

DeepSeek

BLOCKER The peak/off-peak billing flagged last pass went live on schedule, 2026-08-16 at 16:00 UTC. V4-Flash is now $0.22/$0.66 off-peak and $0.44/$1.32 at peak; V4-Pro is $0.66/$1.98 off-peak and $1.32/$3.96 at peak. Peak hours are 01:00–04:00 and 06:00–10:00 UTC, so a nightly batch job scheduled inside one pays double — this is no longer a future risk, it is live. source https://api-docs.deepseek.com/quick_start/pricing

WATCH The margin over the board's other cheap tiers narrowed more than 'still cheapest' implies: V4-Flash's off-peak input ($0.22) is already above GPT-5.6 Luna's flat $0.20, though its off-peak output ($0.66) still undercuts Luna's $1.20. Compare on your own workload's in/out ratio rather than assuming DeepSeek wins on both. source https://api-docs.deepseek.com/quick_start/pricing

BENEFIT The caching discount survives the increase and stays the most aggressive on the board: a cache hit is about 3% of a cache miss off-peak ($0.007 against $0.22), against the 10% that is standard elsewhere. source https://api-docs.deepseek.com/quick_start/pricing

BENEFIT Reported as MIT-licensed with weights on Hugging Face, so self-hosting and fine-tuning carry no commercial restriction. source https://www.marktechpost.com/2026/07/18/kimi-k3-vs-deepseek-v4-pro-vs-glm-5-2-open-trillion-scale-moe-models-compared-on-benchmarks-license-and-serving-cost/

WATCH The pricing page states rates "may vary and DeepSeek reserves the right to adjust them" — and it exercised that on three days' notice for the peak/off-peak change. There is no price-stability commitment to build a margin on; treat every DeepSeek rate as provisional. source https://api-docs.deepseek.com/quick_start/pricing

BENEFIT Alongside the price change, reasoning effort is now documented in three named levels (low, high, max) rather than just 'controllable', and the API added native OpenAI Responses API support with one-click Codex integration. source https://api-docs.deepseek.com/news/news260813

BLOCKER For regulated data, the hosted API's jurisdiction and residency will end the conversation with your compliance team. Self-hosting is the answer here and it is a real one, because the licence permits it. source https://api-docs.deepseek.com/quick_start/pricing

---

Mitigation: Price every DeepSeek workload against the live peak/off-peak tiers, not the pre-2026-08-16 flat rate; move batch work out of 01:00–04:00 and 06:00–10:00 UTC; self-host for regulated or latency-critical paths.

xAI · Grok

BENEFIT grok-4.3 lists a 1M-token window at $1.25 in / $2.50 out below 200K — among the cheapest large-window rates on the board. source https://docs.x.ai/docs/models

WATCH Grok 4.6 (2026-08-12) matches Grok 4.5 on the headline $2/$6 but raises cached input from $0.30 to $0.50 per 1M. On a prefix-heavy agent loop the newer model is the more expensive one — check your cache-hit rate before migrating on the version number. source https://docs.x.ai/docs/models

WATCH Every model doubles at or above 200K input tokens: grok-4.6 goes from $2/$6 to $4/$12. The advertised 500K–1M window costs 2x per token to actually fill. source https://docs.x.ai/docs/models

WATCH The models page states neither max output tokens nor per-model vision support, so size responses against measured behaviour rather than a documented ceiling. source https://docs.x.ai/docs/models

---

Mitigation: Keep prompts under the 200K step, or price the workload at the upper tier from the start.

Moonshot · Kimi

BENEFIT K3 carries a 1,048,576-token window priced flat across the whole thing — no length tiering, unlike Gemini and Grok. source https://platform.kimi.ai/docs/pricing/chat-k3

BENEFIT Cache hits at 10% of the input rate, and a dedicated cheaper coding model (K2.7 Code at $0.95/$4.00, now also documented at a 262,144-token context). source https://platform.kimi.ai/docs/pricing/chat-k27-code

WATCH Weights are reported to ship under a Modified MIT licence with commercial thresholds — a separate agreement above roughly $20M model-as-a-service revenue, and on-screen credit above 100M monthly active users. Fine for most teams; check it if you are either of those. source https://www.marktechpost.com/2026/07/18/kimi-k3-vs-deepseek-v4-pro-vs-glm-5-2-open-trillion-scale-moe-models-compared-on-benchmarks-license-and-serving-cost/

BENEFIT The pricing page that was restructured to only show legacy moonshot-v1 rates last pass now links out to real per-model pages for K3, K2.7 Code and K2.6, and all three are first-party-citable again. K2.6 came back cheaper than the host-sourced rate this board carried: $0.95/$0.16/$4.00, down from $1.20/$0.20/$4.50. source https://platform.kimi.ai/docs/pricing

WATCH moonshot-v1's full platform sunset is 2026-08-31 — not a board row, but worth knowing if you're still pinned to it. source https://platform.kimi.ai/docs/pricing

---

Mitigation: Confirm rates against the first-party console before you sign anything; read the licence if you are at scale.

Z.ai · GLM

BENEFIT GLM-4.7-FlashX at $0.07 in / $0.40 out is the cheapest first-party lab row on the board — only a host undercuts it — and GLM-4.7-Flash is free on the API outright. source https://docs.z.ai/guides/overview/pricing

BENEFIT Cached-input storage is now stated as "limited-time free" across the GLM line — a cost that has no equivalent stated end date, so it's worth re-checking before you build a margin on it staying free. source https://docs.z.ai/guides/overview/pricing

BENEFIT The GLM-5 line is reported as MIT-licensed with full weights on Hugging Face — unrestricted commercial use if you self-host. source https://www.marktechpost.com/2026/07/18/kimi-k3-vs-deepseek-v4-pro-vs-glm-5-2-open-trillion-scale-moe-models-compared-on-benchmarks-license-and-serving-cost/

WATCH The pricing page states no context windows and no max-output limits, so you are sizing prompts against measured behaviour rather than a documented contract. source https://docs.z.ai/guides/overview/pricing

WATCH A thinner SDK and ecosystem than the US labs — budget for writing your own retry, streaming and observability layer. source https://docs.z.ai/guides/overview/pricing

---

Mitigation: Measure the real context limit yourself before designing around it.

Alibaba · Qwen

BENEFIT qwen3.7-max lists a 1M-token window with 65,536 max output, and Model Studio now documents caching as a multiplier rather than a per-model rate: explicit cache hits bill at 10% of the input rate, cache creation at 125%. source https://www.alibabacloud.com/help/en/model-studio/context-cache

WATCH Last week this board called that a flat "90% cached-input discount". That was half right: 10% applies to explicit caches, but implicit caching — the one you get without asking — bills at 20% of the input rate on most models, twice as much. Which of the two you are on decides the number, and they are mutually exclusive per request. source https://www.alibabacloud.com/help/en/model-studio/context-cache

BENEFIT The smaller Qwen lines are Apache-2.0 — the least legally ambiguous open weights on this board for a commercial product. source https://www.marktechpost.com/2026/07/18/kimi-k3-vs-deepseek-v4-pro-vs-glm-5-2-open-trillion-scale-moe-models-compared-on-benchmarks-license-and-serving-cost/

WATCH The Max tier is closed-weight; only the smaller models are open. Do not plan a self-hosting story around Max. source https://www.marktechpost.com/2026/07/18/kimi-k3-vs-deepseek-v4-pro-vs-glm-5-2-open-trillion-scale-moe-models-compared-on-benchmarks-license-and-serving-cost/

WATCH First-party and host prices diverge sharply — around $2.50 per 1M input on Model Studio against $1.25 on Together for the same model. Shop it before you commit. source https://www.together.ai/pricing

---

Mitigation: Price the same model at two hosts before picking one; keep the open lines for anything you intend to fine-tune.

Meta · Llama & Muse

WATCH Meta runs a first-party paid API — the Meta Model API — but only for Muse Spark, its first closed model. source https://dev.meta.ai/docs/pricing-rate-limits.md

BENEFIT Muse Spark bills one flat rate across its 1M window with no long-context premium, at $1.25 in / $0.15 cached / $4.25 out — cheaper per token than every frontier row on this board bar the small models. source https://developer.meta.com/ai/models/muse-spark/

BENEFIT Muse Glimmer 30B (2026-08-10) is Meta's first Apache-2.0 release since the Llama 4 Community Licence backlash — a dense model distilled from Muse Spark. No first-party endpoint serves it; see the hosts table for a price. source https://research.meta.ai/blog/introducing-muse-glimmer-open-agentic-model

WATCH Llama 4 Scout is reported at a 10M-token context window — by a wide margin the largest on this board, open-weight or otherwise — but the secondary source cited for that figure no longer states it when re-checked this pass. Carried forward as the best available citation, not freshly confirmed; see the coverage ledger. source https://presenc.ai/research/open-weight-license-landscape-2026

BLOCKER The split matters more than any single launch. Muse Spark is closed-weight on Meta's own endpoint; Muse Glimmer is open-weight but also with no first-party endpoint; Llama remains open-weight with no first-party endpoint at all. Every Llama and Glimmer price on this board is a host's price, labelled as such in that row's sources. There is no single "Meta price" — see the hosts table. source https://dev.meta.ai/docs/pricing-rate-limits.md

WATCH The Llama 4 Community Licence is not OSI-approved: commercial use is permitted up to 700M monthly active users, with an acceptable-use policy attached. "Open weights" here is not Apache-2.0 — Muse Glimmer is the first Meta row on this board where it is. source https://presenc.ai/research/open-weight-license-landscape-2026

Mitigation: Treat Muse Spark and Llama as two unrelated products. For Llama, pick the host first, then price the workload; for Muse Spark, read Meta's own terms — the Llama licence does not apply.

MiniMax

BENEFIT M3 lists at $0.30 in / $1.20 out with an 80% cached-input discount — one of the better price-to-caching ratios at the cheap end of the board. source https://www.together.ai/pricing

WATCH The advertised 1M window is not flat-priced: input above 512K steps up to $0.60 in / $2.40 out, and the headline rate is a standing promotional 50% discount off that same $0.60/$2.40. Two reasons the cheap number may not be the number you pay. source https://openrouter.ai/minimax/minimax-m3

WATCH The first-party pricing page did not resolve when we checked, so rates here come from a host and a model directory rather than MiniMax itself. Open-weight status is also unverified — treat self-hosting as an open question until you have read the licence.

Mitigation: Verify licence and first-party rates directly before designing around this model.

Perplexity · Sonar

BENEFIT Search-grounded answers with citations as a first-class API, which is otherwise a whole retrieval stack you would build and maintain. source https://docs.perplexity.ai/getting-started/pricing

BLOCKER The per-request search fee usually dominates the token cost: $5–$14 per 1,000 requests depending on search-context size, on top of per-token rates. Model the request fee first or your estimate will be wrong by an order of magnitude — this is the one provider where the token rate is the small number. source https://docs.perplexity.ai/getting-started/pricing

WATCH The pricing page now states that Sonar Chat Completions has become the Agent API, with a migration guide alongside it. The rates are unchanged, but the surface you call them through is moving — read the guide before your next integration. source https://docs.perplexity.ai/getting-started/pricing

---

Mitigation: Estimate cost per request, never per token, and pin the search-context size explicitly.

Groq `inference host`

BENEFIT Throughput an order of magnitude above general-purpose clouds — 1,000 tokens/sec on GPT OSS 20B and 840 on Llama 3.1 8B. If time-to-first-token is the product, this is the lever. source https://groq.com/pricing

WATCH A narrow menu: you get the open-weight models they have deployed, on their schedule. No frontier closed models. source https://groq.com/pricing

WATCH Context caps sit at 128K–131K across the menu, well below the 1M windows available elsewhere, and the caching discount applies only to select models. source https://groq.com/pricing

WATCH Groq's pricing page stopped publishing a per-model rate table this pass, and the console pricing path 404s. The rates in the hosts table are carried forward from 2026-08-06 unverified — confirm in the console before you commit. source https://groq.com/pricing

---

Mitigation: Use it for the latency-critical hop; keep a large-window provider for the rest.

Together AI `inference host`

BENEFIT The broadest open-weight menu on one bill — Llama, Qwen, DeepSeek, Kimi, GLM, MiniMax and now Meta's Muse Glimmer — with published cached-input rates. source https://www.together.ai/pricing

BLOCKER Checkpoints can carry two labels for what looks like the same model at different prices. This pass, Together lists both a legacy 'DeepSeek V4 Pro' ($1.74/$0.20/$3.48, the old markup) and a new 'DeepSeek V4 Pro 0813' ($1.32/$0.13/$3.96, which matches DeepSeek's own new peak-hour rate on in/out) side by side. Pin the exact model string, not just the family name, or you won't know which one you're billed at. source https://www.together.ai/pricing

---

Mitigation: Always price the model first-party before accepting a host's rate.

Hugging Face `hub`

BENEFIT The one place where weights, licence, provenance and eval history sit together. For open-weight due diligence there is no substitute, and routed inference is charged at the provider's own rate with no Hugging Face markup. source https://huggingface.co/docs/inference-providers/pricing

BLOCKER A hub is not an endpoint. Included credits are $0.10/month for free accounts and $2.00 for PRO, and `hf-inference` itself now focuses on CPU inference and smaller historical models — so treating it as a production API surprises people. source https://huggingface.co/docs/inference-providers/pricing

WATCH Because requests route to whichever provider backs the model, availability, limits and price are that provider's to change, not Hugging Face's. source https://huggingface.co/docs/inference-providers/pricing

---

Mitigation: Use it to choose and audit a model; deploy where you control the endpoint.

WHERE TO RUN IT

Inference hosts price other people's models.

They belong in their own table: a host doesn't determine what a model can do, it determines what it costs you and how fast it arrives. The markups are real and rarely advertised — the same open weights can differ by 4× between a host and the lab that made them.

GROQ

VERIFIED 2026-08-27

Speed is the product — up to 1,000 tokens/sec on a narrow menu, at a 131,072-token context cap. Per-model rates are citable again this pass: groq.com/pricing still publishes no rate table, but console.groq.com/docs/models now carries prices, contexts and speeds, and Llama 3.3 70B Versatile is back on that page after vanishing from it last pass. The catch is that both Llama rows now read "ContactSales" instead of a rate — their prices below are carried forward from 2026-08-06 and are no longer publicly quoted.

MODEL	IN	CACHED IN	OUT	CONTEXT	THROUGHPUT	NOTE
GPT OSS 20B	$0.075	—	$0.3	131K	1000 tok/s	fastest row published; context and speed re-read this pass
GPT OSS 120B	$0.15	—	$0.6	131K	500 tok/s	
Llama 3.1 8B Instant	$0.05	—	$0.08	131K	560 tok/s	cheapest row on the board — but the price is carried forward from 2026-08-06: this model now reads "ContactSales" on Groq's models page. Context and the 560 tok/s were re-read this pass (down from the 840 tok/s recorded on 2026-08-06)
Llama 3.3 70B Versatile	$0.59	—	$0.79	131K	280 tok/s	back on Groq's models page after disappearing from it last pass, so availability is confirmed again — but priced "ContactSales", so $0.59/$0.79 is carried forward from 2026-08-06, not current. Together lists this model at $1.04/$1.04. Speed re-read this pass at 280 tok/s, down from 394
Qwen 3.6 27B	$0.6	—	$3.00	131K	500 tok/s	preview model; context corrected to 131,072 from the 131,000 this board carried

TOGETHER AI

VERIFIED 2026-08-27

The broadest open-weight menu on one bill — but check the markup against first-party rates before you commit, and check which checkpoint label you're actually being routed to.

MODEL	IN	CACHED IN	OUT	CONTEXT	THROUGHPUT	NOTE
DeepSeek V4 Pro 0813	$1.32	$0.13	$3.96	—	—	matches DeepSeek's own new peak-hour rate on in/out exactly, but at 3x DeepSeek's peak cached-hit rate ($0.044). Together separately still lists a legacy 'DeepSeek V4 Pro' row unchanged at $1.74/$0.20/$3.48 — which checkpoint you're routed to needs checking before you price a workload
DeepSeek V4 Flash 0731	$0.14	$0.03	$0.28	—	—	the pre-2026-08-16 flat rate, still served here under a dated checkpoint label — half DeepSeek's own current off-peak input and a fifth of its peak. A cheap row that is cheap because it is an older checkpoint, which is the trap this table exists to show
Kimi K3	$3.00	$0.3	$15.00	—	—	matches first-party (platform.kimi.ai)
Kimi K2.7 Code	$0.95	$0.19	$4.00	—	—	matches first-party (platform.kimi.ai)
Qwen3.7-Max	$1.25	$0.13	$3.75	—	—	half the first-party Model Studio rate
GLM-5.2	$1.40	$0.26	$4.40	—	—	matches first-party
MiniMax M3	$0.3	$0.06	$1.20	—	—	
Llama 3.3 70B	$1.04	—	$1.04	—	—	vs $0.59/$0.79 last recorded on Groq, where this model no longer appears at all
Muse Glimmer 30B	$0.35	$0.04	$1.50	—	—	Meta's first Apache-2.0 model; no first-party endpoint to compare against

HUGGING FACE **VERIFIED 2026-08-27**

A hub, not an endpoint. Routed inference is charged at the provider's own rate with no Hugging Face markup — but the included credits are pocket change and hf-inference now targets CPU-scale models.

MODEL	IN	CACHED IN	OUT	CONTEXT	THROUGHPUT	NOTE
Routed to any integrated provider	–	–	–	–	–	provider's own rate, passed through with no markup
Included credits — Free account	–	–	–	–	–	$0.10 per month
Included credits — PRO	–	–	–	–	–	$2.00 per month
Included credits — Team / Enterprise	–	–	–	–	–	$2.00 per seat per month, pooled
hf-inference	–	–	–	–	–	billed by compute time × hardware rate; now mostly CPU and smaller models

SINCE WE STARTED

What has moved, and who moved it.

Every refresh is kept, so a price on this board has a history rather than just a value. The distinction that matters: a **vendor change** is the provider changing what it charges. A **board correction** is us changing what we print — the provider charged the same before and after. Plotting the second as the first would publish an increase that never happened, so the two are labelled and never merged.

DeepSeek-V4-Flash
deepseek

07–30 **$0.28** → 08–06 **$0.28** → 08–14 **$0.28** → 08–20 **$0.66** → 08–27 **$1.32** → 09–03 **$1.32**

VENDOR CHANGE 2026-08-20 · in, cached_in, out — The peak/off-peak change announced 2026-08-13 went live on schedule at 2026-08-16 16:00 UTC. At this pass the board printed the off-peak rate — see the 2026-08-27 correction.

OUR CORRECTION 2026-08-27 · in, cached_in, out — Re-based from the off-peak rate to the peak rate, which DeepSeek's own pages call the standard rate. DeepSeek charged the same before and after; the board's column header promises list rates and was printing a time-of-day discount.

DeepSeek-V4-Pro
deepseek

07–30 **$0.87** → 08–06 **$0.87** → 08–14 **$0.87** → 08–20 **$1.98** → 08–27 **$3.96** → 09–03 **$3.96**

VENDOR CHANGE 2026-08-20 · in, cached_in, out — As above: live 2026-08-16, printed here at the off-peak rate.

OUR CORRECTION 2026-08-27 · in, cached_in, out — Re-based from off-peak to peak on the same grounds. No change to what DeepSeek charges.

Gemini 3.6 Flash
google

07–30 **$7.50** → 08–06 **$7.50** → 08–14 **$3.75** → 08–20 **$3.75** → 08–27 **$3.75** → 09–03 **$3.75**

VENDOR CHANGE 2026-08-14 · in, cached_in, out — Halved to $0.75/$0.075/$3.75 when Gemini 3.7 Flash superseded it three weeks after launch.

GPT-5.6 Luna
openai

07–30 **$6.00** → 08–06 **$1.20** → 08–14 **$1.20** → 08–20 **$1.20** → 08–27 **$1.20** → 09–03 **$1.20**

VENDOR CHANGE 2026-08-06 · in, cached_in, out — OpenAI cut Luna by 80%, effective 2026-07-30: $1.00/$0.10/$6.00 to $0.20/$0.02/$1.20.

OUR CORRECTION 2026-08-06 · cache_write — The pricing page had stated this rate all along and the board carried it as null. OpenAI disclosed nothing new; we filled a gap of our own making.

GPT-5.6 Sol
openai

07–30 **$30.00** → 08–06 **$30.00** → 08–14 **$30.00** → 08–20 **$30.00** → 08–27 **$20.00** → 09–03 **$20.00**

OUR CORRECTION 2026-08-06 · cache_write — As above — the page stated the cache-write rate and the board had carried null.

VENDOR CHANGE 2026-08-27 · in, cached_in, cache_write, out — OpenAI cut Sol from $5.00/$0.50/$6.25/$30.00 to $4.00/$0.40/$5.00/$20.00. The pricing page carries no announcement date, so this is recorded on the date the board observed it.

GPT-5.6 Terra
openai

07–30 **$15.00** → 08–06 **$12.00** → 08–14 **$12.00** → 08–20 **$12.00** → 08–27 **$12.00** → 09–03 **$12.00**

VENDOR CHANGE 2026-08-06 · in, cached_in, out — OpenAI cut Terra by 20%, effective 2026-07-30: $2.50/$0.25/$15.00 to $2.00/$0.20/$12.00.

OUR CORRECTION 2026-08-06 · cache_write — As above — the page stated the cache-write rate and the board had carried null.

Kimi K2.6
moonshot

07–30 **$4.50** → 08–06 **$4.50** → 08–14 **$4.50** → 08–20 **$4.00** → 08–27 **$4.00** → 09–03 **$4.00**

OUR CORRECTION 2026-08-20 · in, cached_in, out — Moonshot's per-model pricing pages became citable again, and the first-party rate was lower than the host figure the board had been carrying. Moonshot did not cut the price; we stopped quoting a host for it.

Qwen3.5-397B-A17B
alibaba

07–30 **$3.60** → 08–06 **$3.60** → 08–14 **$3.60** → 08–20 **$3.60** → 08–27 **$3.60** → 09–03 **$3.60**

OUR CORRECTION 2026-08-06 · cached_in — Withdrawn on the same grounds as Qwen3.7-Max — the $0.35 was a host's number.

Qwen3.7-Max
alibaba

07–30 **$7.50** → 08–06 **$7.50** → 08–14 **$7.50** → 08–20 **$7.50** → 08–27 **$7.50** → 09–03 **$7.50**

OUR CORRECTION 2026-08-06 · cached_in — Withdrawn, not cut: the $0.25 cached rate came from a host, and Alibaba's own page states no cached-input rate. A null here means unstated, not free.

Qwen3.7-Plus
alibaba

07–30 **$1.28** → 08–06 **$1.60** → 08–14 **$1.60** → 08–20 **$1.60** → 08–27 **$1.60** → 09–03 **$1.60**

OUR CORRECTION 2026-08-06 · in, out — Re-sourced from Together AI to Alibaba's own Model Studio page. The first-party rate is higher than the host rate the board had been printing; Alibaba did not raise anything.

Series data: `/data/history/index.json`. Why each point moved, in our own words, is in `changes.json` — including the ones that were our fault.

THIS WEEK IN MODELS

What changed, and whether you should care.

Every week. The second line is the one that matters: what it changes for someone shipping. This week's full write-up: Nobody changed a price this week.

2026-09-03

Gemini 3.8 Flash arrives at Gemini 3.7 Flash's exact price

Google's models page marks it "New Stable" and its pricing page gives $0.75 in / $0.075 cached / $3.75 out with a 1M window and 64K max output — column for column, what 3.7 Flash costs. A free upgrade, but it inherits the clock rather than resetting it: the same page says both rates double to $1.50/$7.50 on 2027-01-01, and cache storage goes $0.50 to $1.00 per 1M/hr on the same day. Thinking is tunable low/medium/high and defaults to medium, so a cost model measured on low will not match the bill. source https://ai.google.dev/gemini-api/docs/pricing

---

2026-09-03

Claude Fable 5.1 and Mythos 5.1 price cache reads at 0.025x, not 0.1x

Same $10 in / $50 out as Fable 5, but the cache-hit rate is $0.25 against Fable 5's $1.00 — Anthropic's pricing page states these two rows use a 0.025x multiplier where every other Claude model uses 0.1x. On a workload that re-reads a large fixed prefix, the 5.1 rows are four times cheaper to cache than the 5.0 rows they replace, at an identical headline price. Both keep the 512-token minimum prefix. Mythos 5.1 is limited-availability, so the rate is published but not simply purchasable. source https://platform.claude.com/docs/en/about-claude/pricing

---

2026-09-03

GLM-5.3's weights are open, and the licence is not MIT

The Hugging Face card for GLM-5.3 tags the licence "glm-5.3" — a bespoke licence — where GLM-5.2 carries MIT. The card does not print the terms and docs.z.ai states no licence at all, so what it permits is not something this board can tell you; what it can tell you is that the MIT assumption people carry across the GLM line does not hold here. If your open-weights requirement is contractual rather than aesthetic, this is a row to read before you pull. The API price is unchanged from GLM-5.2 at $1.4/$4.4. source https://huggingface.co/zai-org/GLM-5.3

2026-09-03

GPT-5.6 Sol's $4.00 / $20.00 is now labelled promotional, through at least 2026-11-21

The rate did not move this pass — the wording did. OpenAI's pricing page now calls the short-context rate promotional and says it is available "at least through November 21, 2026". Last pass it was simply the price. Nothing is scheduled to change and "at least" is doing real work in that sentence, but a row that reads promotional is a row to budget at its pre-cut $5.00/$30.00 rather than assume permanence. The long-context tier's trigger point is still undocumented for the 5.6 series. source https://developers.openai.com/api/docs/pricing

---

2026-09-03

Perplexity's Sonar Chat Completions endpoint is supported only until 2026-09-27

The pricing page now reads "Sonar Chat Completions is now Agent API" and "Sonar will be supported until September 27, 2026" — twenty-four days from this refresh, and inside the window of the next two. Read it precisely: the date is on the endpoint, not on the models. No Sonar model is marked deprecated and all three keep their rates. But if you call Sonar through Chat Completions, you have a migration to schedule and less than a month to schedule it in. `source` https://docs.perplexity.ai/getting-started/pricing

METHOD & CORRECTIONS

How this board is built, and how to prove us wrong.

UNITS

USD per 1,000,000 tokens. **List rate** — no committed-use, enterprise or volume discount. Batch, off-peak and long-context surcharges are shown separately.

SOURCING

Every row records its source URLs and the date they were fetched. A validator refuses to publish a model with no source, or one whose check is more than 14 days old — so the board cannot quietly rot.

MISSING NUMBERS

Where a provider's own page doesn't state something, we print `?` or leave it out of the charts and say why. **We never fill a gap with a plausible number.** The validator enforces this: a null value without an explanation fails the build.

NO BENCHMARKS HERE

Deliberately. We have not independently evaluated these third-party models, and quoting vendors' own scores as if we had would be exactly the thing we tell clients not to do. Our own models ship with their own eval gates.

JUDGEMENT COLUMNS

"Best for" and "Watch out" are engineering judgement from building on these APIs. They're opinion, dated and revisable — everything factual next to them is linked.

MACHINE-READABLE

The same data, versioned: `/data/model-board.json`. No key, no rate limit, CORS open, free to use with attribution — the terms, the schema and the rules for quoting a number are on the data page.

COVERAGE LEDGER — WHAT THIS PASS DID NOT COVER

- **Groq — Llama per-model rates** — unchanged in substance, confirmed again this pass. console.groq.com/docs/models lists Llama 3.3 70B Versatile with its context (131,072) and max completion (32,768) but prints "ContactSales" where a rate would go, so the price stays carried from 2026-08-06 and flagged on the row. groq.com/pricing is still a marketing page with no rate table at all
- **Cohere** — the public pricing page lists only legacy Command models; current-generation rates were not on a page we could source this pass
- **Amazon Nova / Bedrock** — per-model rates are spread across a region-keyed pricing console rather than a citable table
- **Google Vertex AI** — same as Bedrock — regional pricing, no single citable table
- **Llama 4 Scout's 10M-token context claim** — third pass running that this could not be re-sourced. presenc.ai/research/open-weight-license-landscape-2026 was re-fetched today and is still dated 'Last updated: May 2026'; it still states the 700M-MAU licence threshold and the name-attribution rules, and it still does not state the 10M context figure. The row keeps the value from the last pass that confirmed it, and now says so in its own note. The licence half of the citation verifies cleanly, which is why verified.on moves this pass while the window stays flagged
- **Fireworks AI** — not fetched this pass
- **OpenRouter** — the models list renders client-side, so no server-side table to cite; individual model pages are cited where a row needs them
- **Z.ai GLM-5.1 and GLM-5-Turbo** — both still on the pricing page this pass at $1.4/$0.26/$4.4 and $1.2/$0.24/$4.0, unchanged. Not added — the board carries a curated set per lab and GLM now has five rows with GLM-5.3. Recorded here so the omission reads as a choice rather than an oversight
- **Nvidia Nemotron, Microsoft Phi** — not fetched this pass
- **xAI Grok 4.20** — found on xAI's models page this pass, priced identically to Grok 4.3. It is not a new release — GA since March 2026, predating 4.3/4.5/4.6 despite the higher version number — so it isn't a news item, but it has never been on this board. Left off pending a decision on whether the curated set should include it
- **Context windows for several OpenAI, Google, Mistral and Z.ai rows** — the sourced pricing pages do not state them; shown as 'not stated' rather than guessed. Three exceptions this pass, all newly filled from first-party pages: Gemini 3.8 Flash (1M / 64K, latest-model page), GLM-5.3 (1M / 128K, docs.z.ai) and Claude Haiku 4.5 (200K / 64K, models overview) — the last of which had been null since the row was added
- **The gpt-5.6 long-context threshold** — re-checked again this pass and still open: OpenAI prints a full long-context column for Sol, Terra and Luna ($8.00/$0.80/$10.00/$30.00 on Sol) but names a 272K threshold only for the gpt-5.5 and gpt-5.4 rows, so the 5.6 tier remains priced but untriggered on paper. New this pass, on the same page: the $4.00/$20.00 short-context rate is now labelled promotional "at least through November 21, 2026"
- **Mistral — caching mechanics** — opened this pass as the narrower remainder of the cached-input item now closed. docs.mistral.ai/inference/pricing prints a per-model cached rate, so the four live Mistral rows finally carry one, but no Mistral page this board cites says whether caching is automatic or explicit, gives a TTL, or names a minimum cacheable prefix. caps.caching.kind stays 'unverified' on every Mistral row: knowing the rate is not knowing the mechanism, and the 2026-08-29 correction on this same provider came from filling exactly that kind of gap by inference
- **GLM-5.3's licence terms** — the Hugging Face card tags the weights 'glm-5.3' — demonstrably not the MIT that GLM-5.2 carries — but does not print the licence text, and docs.z.ai does not state a licence at all. Secondary reporting describes a revenue-gated condition on large Model-as-a-Service operators; that is not sourced to a page this board can cite, so the row states only what the card states: bespoke, and not MIT
- **Gemini 3.5 Flash-Lite** — on both Google's models list and its pricing page this pass, but the rates as fetched read identically to Gemini 2.5 Flash ($0.30/$0.03/$2.50), which is more likely a fetch artefact than a fact. Not added on a reading that ambiguous — a row is worth less than nothing if its price came from a page we misread
- **Alibaba qwen3.8-flash** — appears on alibabacloud.com/help/en/model-studio/models this pass alongside qwen3.8-max and qwen3.7-plus, with no rate on that page. Not added pending a citable price; recorded here so the omission reads as a choice

Found a wrong number? Send the source URL to contact@naderu.com and it goes into the next weekly update, with a changelog line saying what was wrong.

▼ CHANGELOG

2026-09-03

- Added Gemini 3.8 Flash at $0.75/$0.075/$3.75, 1M window and 64K max output — identical to Gemini 3.7 Flash in every price column, on the same introductory timetable ending 2026-12-31. Gemini 3.7 Flash moves to previous.
- Added Claude Fable 5.1 and Claude Mythos 5.1 at $10.00/$0.25/$12.50/$50.00. The cache-read rate is the difference that matters: 0.025x base against the 0.1x every other Claude row pays, so $0.25 where Fable 5 charges $1.00. Claude Fable 5 and Claude Mythos 5 move to previous, both now listed under Legacy on Anthropic's own overview page.
- Added GLM-5.3 at $1.4/$0.26/$4.4 with a first-party 1M window and 128K max output from docs.z.ai — GLM-5.2's window is still only second-hand. Weights are open but the Hugging Face card tags the licence "glm-5.3", not MIT; the card does not print the terms, so the row says bespoke and not MIT and nothing further. GLM-5.2 moves to previous.
- Added Muse Spark 1.3 at $1.25/$0.15/$4.25, identical to 1.2 and 1.1, which both remain on the pricing page. Muse Spark 1.2 moves to previous.
- RESOLVED — Mistral per-model cached input, deferred since 2026-08-29 and carried forward explicitly for this pass. docs.mistral.ai/inference/pricing does print the column filled per model, and it is now cited by all four live Mistral rows: Medium 3.5 $0.15, Small 4 $0.015, Codestral $0.03, Large 3 $0.05. Devstral 2 is absent from that page, consistent with its retirement, and stays null. The −90% the other two pages described only in prose is real, and it took a third page to source it.
- Mistral caching MECHANICS stay unverified, and the deferred ledger now says so as its own entry. A rate is not a mechanism: no Mistral page this board cites states whether caching is automatic or explicit, gives a TTL, or names a minimum prefix. caps.caching.kind stays "unverified" on every Mistral row. Filling that gap by inference is precisely what produced the 2026-08-29 correction on this same provider.
- gpt-5.6-sol: the $4.00/$20.00 rate is unchanged but is now labelled promotional "at least through November 21, 2026" on OpenAI's pricing page — new wording this pass, recorded on the row and in news.
- Perplexity: all three Sonar rows gain a note that the Chat Completions endpoint is supported only until 2026-09-27. The models are NOT deprecated and their generation is unchanged — the date is on the endpoint, and this board checked that distinction before moving anything.
- Filled from first-party pages, all previously null: Claude Opus 5 and Claude Sonnet 5 max output 128,000; Claude Haiku 4.5 window 200,000 and max output 64,000, plus its retirement commitment of no sooner than 2026-10-15, the nearest date on this board; MiniMax M3 max output 262,144.
- Muse Glimmer 30B: reasoning moves from unverified to controllable and tools from null to true, both stated on the Hugging Face card and neither previously read off it.
- llama-3.3-70b: Llama 3.3 70B Versatile is on console.groq.com/docs/models again with its context (131,072) and max completion (32,768), so its availability there is confirmed rather than merely assumed — but the row reads "ContactSales", so Groq still publishes no rate and the price stays carried from 2026-08-06.
- llama-4-scout: the cited source was re-fetched and re-confirms the 700M-MAU licence terms and the name-attribution rules, but for the third pass running does not state the 10M context figure. verified.on moves to 2026-09-03 on the strength of the licence half; the window keeps its flag, and the row now says which half verified.
- Job vocabulary, pass 2 of 3 (naderu#304): summarisation, translation and copy-editing ship with their cards, taking jobs[] from four slugs to seven of fourteen. code-completion, code-review, deep-reasoning, rag-grounded-answers and the media jobs remain for the 09-10 pass. constraints[] and postures[] are unchanged and complete.
- Re-read constraints[] and postures[] against this pass's data rather than last pass's, which the gates cannot check and which had gone stale in two places that mattered. The open-weights card led on GLM-5.3-Flash and Codestral as its examples, both added on 2026-08-27 but described as "added this pass"; it now leads on GLM-5.3, whose bespoke non-MIT licence sitting directly above GLM-5.2's plain MIT is a sharper version of the same point. The quality posture still named Fable 5 and Mythos 5 as the top of the board and has moved to the 5.1 rows, including the 0.025x cache read that makes the newer row cheaper at an identical headline price. Four further cards — long-context, low-latency, no-training-on-data, saver and balanced — had prose dating earlier passes' events to this one.
- New in the deferred ledger: Gemini 3.5 Flash-Lite (on both Google pages, but the rates as fetched read identically to Gemini 2.5 Flash, which is more likely a misread than a fact — not added on a reading that ambiguous), Alibaba qwen3.8-flash (listed with no citable rate) and GLM-5.3's licence terms (tagged but not printed).

- Re-verified unchanged and re-dated: all five OpenAI rows, the five carried-over Anthropic rows, the five carried-over Gemini rows, both DeepSeek rows, all four Grok rows, all three Kimi rows, the four carried-over Z.ai rows, all four Qwen rows including Qwen3.8-Max's $0.25 cached rate confirmed against its own model page, Muse Glimmer 30B, MiniMax M3 at Together's unchanged $0.30/$0.06/$1.20, and all three Sonar rows.

2026-08-29

- CORRECTION — Mistral prompt caching. This board carried caps.caching.kind: "none" on Mistral Medium 3.5, Small 4, Large 3 and Devstral 2, a Mistral blocker note reading "No prompt caching is documented on the API pricing page. A long fixed system prompt is billed in full on every single call", and a watch_out on Medium 3.5 repeating it. None of that was on a sourced page — it was inferred from the absence of a cached-input column on mistral.ai/pricing/api/. The four rows now read "unverified", the blocker is downgraded to a watch, and the invented consequence is removed. No price on any Mistral row changed; what changed is a claim this board made and could not source. Reported as naderu#320.
- The 2026-08-14 pass did not miss this — it reached the wrong conclusion and wrote it down. That pass re-checked the –90% line, confirmed by direct fetch that it was generic "Configure your API" copy with "no cached-input figure ... tied to Medium 3.5, Small 4 or Large 3", and reaffirmed the blocker on that basis. True of the page it read, false of the provider: docs.mistral.ai/inference/pricing prints a Cached input column filled per model at 10% of input — Mistral Large 3 at $0.05 against $0.50 input, and the same ratio down the table — read on 2026-08-29. It is a third Mistral pricing page, cited by neither the rows nor the provider entry, so no refresh ever had it in its work queue.
- Those per-model rates are not back-dated into this pass. They land with the 2026-09-03 refresh carrying verified.on: 2026-09-03 and docs.mistral.ai/inference/pricing as a source, because a rate read on the 29th cannot honestly ship under a board dated the 27th. Until then the gap is in coverage.deferred, and this entry carries the figure with the date it was read.

2026-08-27

- OpenAI: GPT-5.6 Sol cut to $4.00/$0.40/$5.00/$20.00 from $5.00/$0.50/$6.25/$30.00; its long-context tier moved to $8.00/$0.80/$10.00/$30.00 from $10.00/$1.00/$12.50/$45.00. Terra, Luna, GPT-5.5 and GPT-5.4-mini re-verified unchanged. The 272K threshold is still named only for the 5.5 and 5.4 rows, not the 5.6 series.
- CORRECTION — DeepSeek is now carried at its peak rate, which its own pages call the standard rate, rather than the off-peak rate this board printed for two passes. V4-Flash $0.22/$0.007/$0.66 -> $0.44/$0.014/$1.32 and V4-Pro $0.66/$0.022/$1.98 -> $1.32/$0.044/$3.96 in the price column. DeepSeek changed nothing this pass; the board was printing a time-of-day discount under a column header that says list rate. Off-peak rates are in each row's note.
- Groq: rates are citable again after two passes deferred. console.groq.com/docs/models publishes prices, contexts and speeds, and Llama 3.3 70B Versatile is back on the page after disappearing last pass. GPT OSS 20B, GPT OSS 120B and Qwen 3.6 27B re-verified at their carried rates; every row's context corrected to 131,072 from 128,000 (131,000 on the Qwen row). Both Llama rows now read "ContactSales", so their prices stay carried from 2026-08-06 and are flagged row by row. Speeds fell where restated: Llama 3.1 8B Instant 840 -> 560 tok/s, Llama 3.3 70B Versatile 394 -> 280 tok/s. groq.com/pricing still has no rate table.
- Added GLM-5.3-Flash at $0.15/$0.03/$0.50 list, half that until 2026-09-09 under a launch promotion. Weights position is not stated on the pricing page, so it is carried as unverified rather than assumed to inherit the GLM line's MIT terms.
- Added Codestral (v25.08) at $0.30/$0.90 — the board tracked agentic coding but had nowhere to put a code-completion model, which is a different capability (FIM), a different latency budget and a different purchase. Listed under Mistral's proprietary Premier tier despite the open-weight history of the name.
- CORRECTION — Devstral 2's retirement date was wrong: docs.mistral.ai states deprecated 2026-05-22 and retired 2026-07-31, not the 2026-06-30 this board carried. The row keeps its last recorded $0.40/$2.00, which is still unverifiable because the model is off the pricing page entirely.
- Restructured the 'pick by the job' vocabulary into jobs, constraints and postures, each entry now carrying a published id slug (naderu#304, pass 1 of 3). The eight old cards mixed jobs with request shapes and deployment constraints, which is why the list could never be finished. Three job cards carry over — agentic-coding, extraction (renamed from 'bulk extraction': volume is a throughput constraint, not a job) and document-understanding — chat-assistant is added, and the five constraint-shaped cards move to constraints and postures with their researched prose intact. The remaining job slugs ship with their cards in the 09-03 and 09-10 passes; an empty job is worse than an absent one.
- Alibaba: explicit-cache mechanics newly documented and filled in — write 125% of input, read 10%, minimum 1,024 tokens, with implicit cache reading at 20% and the two mutually exclusive.
- Together AI: added a 'DeepSeek V4 Flash 0731' row at $0.14/$0.03/$0.28 — the pre-2026-08-16 flat rate, still served under a dated checkpoint label at a fifth of DeepSeek's own peak input.
- Meta: Together AI dropped Llama 4 Scout from its serverless inference table this pass; it now appears there only under fine-tuning, at $3.00 per 1M tokens for LoRA SFT with a $6.00 minimum. Noted on the row, which has no first-party price to begin with.
- Re-verified unchanged and re-dated: all seven Anthropic rows (including the per-model cache minimums), all six Gemini rows, all four Grok rows, all three Kimi rows, Mistral Medium 3.5 / Small 4 / Large 3, all four Qwen rows, Muse Spark 1.2, Muse Glimmer 30B, Llama 3.3 70B, MiniMax M3 and all three Sonar rows.

- Still deferred: Llama 4 Scout's 10M-token context claim — the cited secondary source again does not state the figure when re-fetched, and verified.on stays at 2026-08-14, which puts the row one pass away from tripping the staleness gate. Cohere, Amazon Nova / Bedrock and Google Vertex AI remain unsourceable from a single citable table.
- CORRECTION to this same pass — the four pre-existing Z.ai rows (GLM-5.2, GLM-4.7-FlashX, GLM-5, GLM-4.7) were re-fetched from docs.z.ai and confirmed unchanged, but their verified.on was left at 2026-08-20: the refresh script's bump list omitted the provider. Corrected to 2026-08-27, which is when the page was actually read. Found by the new price-history tool, which flags any row whose verification date lags the refresh; the 14-day staleness gate could not see it because four days is not stale.

2026-08-20

- DeepSeek: the peak/off-peak billing flagged last pass went live on schedule 2026-08-16 16:00 UTC. V4-Flash $0.14/$0.0028/$0.28 → $0.22/$0.007/$0.66 off-peak, $0.44/$0.014/$1.32 at peak (01:00–04:00 and 06:00–10:00 UTC). V4-Pro $0.435/$0.003625/$0.87 → $0.66/$0.022/$1.98 off-peak, $1.32/$0.044/$3.96 at peak. Both rows also gained named reasoning-effort levels (low/high/max) and native OpenAI Responses API support with Codex integration.
- Moonshot: the pricing page that was restructured to show only legacy moonshot-v1 rates is fixed — K3, K2.7 Code and K2.6 are first-party citable again via per-model pages. K2.6 came back cheaper than the host-sourced figure this board carried: $1.20/$0.20/$4.50 → $0.95/$0.16/$4.00. K2.7 Code and K2.6 both gained a documented context window, 262,144, previously unstated.
- Added Qwen3.8-Max, released first-party 2026-08-12, at $2.00/$0.25/$2.50/$6.00 with a 1M-token API window and open (text-only) weights on Hugging Face under a named custom licence. Ends a 3-pass coverage deferral. Qwen3.7-Max demoted to previous.
- Added Muse Glimmer 30B, Meta's first Apache-2.0 release since the Llama 4 Community Licence, released 2026-08-10 and distilled from Muse Spark. No first-party endpoint; priced from Together AI at $0.35/$0.04/$1.50.
- Added Sonar Reasoning Pro (Perplexity) at $2.00/$8.00, sharing Sonar Pro's request-fee schedule.
- CORRECTION-IN-PROGRESS — Llama 4 Scout's 10M-token context claim: the cited secondary source no longer states this figure when re-fetched. Value kept from the last pass it was confirmed rather than dropped or silently re-dated; verified.on left at 2026-08-14 and the gap logged in coverage.deferred.
- Groq: Llama 3.3 70B Versatile no longer appears on Groq's models page at all, on top of the rate table still being down since 2026-08-06. Its continued availability there — not just its price — is now unconfirmed; the hosts-table row and the board's llama-3.3-70b row both flag this.
- Together AI: DeepSeek V4 Pro's listing split into two rows this pass — a legacy 'DeepSeek V4 Pro' still at $1.74/$0.20/$3.48, and a new 'DeepSeek V4 Pro 0813' at $1.32/$0.13/$3.96 that matches DeepSeek's own peak-hour rate on in/out but not on cached input. Board row and hosts table both now flag the checkpoint-label ambiguity.
- Mistral: re-checked the "–90% cached input" line that now appears on the pricing page. Confirmed by direct fetch that it is generic "Configure your API" marketing copy, not a per-model rate — no cached-input figure is tied to Medium 3.5, Small 4 or Large 3. The existing "no prompt caching documented" blocker note stands; no value changed.
- Z.ai: added a note that cached-input storage is now stated as "limited-time free" across the GLM line — worth re-checking before assuming it stays free.
- MiniMax M3: OpenRouter now also states a $0.05 cache-read rate, not previously on the board. Together's canonical $0.30/$0.06/$1.20 is unchanged.
- Rewrote the Bulk extraction, Long context, Cheapest large window, Vision & documents and On-prem/fine-tunable use-case cards — the DeepSeek repricing going live, Llama 4 Scout's shaky citation, and the Qwen3.8-Max/Muse Glimmer additions all made at least one card's wording stale or its model list due for a swap.
- All other rows re-fetched and confirmed unchanged: OpenAI 5, Anthropic 7, Google 6, Mistral 4, xAI 4, Z.ai 4, Alibaba 2 (Plus, the previous-gen 397B), Perplexity 2, MiniMax 0 (note-only), Llama 1 (Llama 4 Scout, value unconfirmed — see above), Muse Spark 1, plus the Together and Hugging Face host blocks.

2026-08-14

- DeepSeek: no rate changed today, but both V4 rows now carry the increase landing 2026-08-16 at 16:00 UTC — V4-Flash $0.14/$0.28 becomes $0.22/$0.66 off-peak and $0.44/$1.32 at peak; V4-Pro $0.435/$0.87 becomes $0.66/$1.98 and $1.32/$3.96. Peak hours are 01:00–04:00 and 06:00–10:00 UTC. This board's own window runs to 2026-08-21, so the rates shown are correct for two of the next seven days and the rows say so.
- Added Gemini 3.7 Flash, released 2026-08-13, at $0.75/$0.075/$3.75 with a 1M window and 64K max output.
- Gemini 3.6 Flash: $1.50/$0.15/$7.50 → $0.75/$0.075/$3.75, and demoted to previous — superseded three weeks after launch.
- Added Grok 4.6, released 2026-08-12, at $2.00/$0.50/$6.00 below 200K; Grok 4.5 demoted to previous. Same headline rate, but cached input rose from $0.30 to $0.50.
- Claude Sonnet 5: the increase to $3/$15 scheduled for 2026-09-01 has been withdrawn. $2/$10 is now the standard rate, not introductory pricing.

- CORRECTION — Meta: this board stated Meta had no first-party API. It does. Added Muse Spark 1.2 at $1.25/$0.15/$4.25 on the Meta Model API, Meta's first closed model, and rewrote the provider row to separate it from Llama, which still has no first-party endpoint.
- CORRECTION — Devstral 2 was retired on 2026-06-30 and this board carried it as a current model for two passes. Moved to deprecated.
- CORRECTION — Alibaba's cached-input discount is not a flat 90%. Explicit cache hits bill at 10% of input, implicit caching at 20% on most models. Last week's note stated only the first, which understated the cost of the caching you get by default.
- Anthropic: added a note that Fable 5 was suspended on 2026-06-12 under export controls and restored on 2026-07-01 — an availability risk that has already happened once.
- Mistral: added a note that Mistral now resells Z.ai's GLM 5.2 on its own API, at a cached rate below Z.ai's own.
- Llama 3.3 70B: max output filled in at 32,768 from Groq's models page.
- MiniMax M3: OpenRouter's listing moved $0.24 → $0.23 input and dropped its discount label; the board's Together-sourced $0.30/$1.20 is unchanged.
- NOT verified this pass: Moonshot's three K-series rows (the pricing page was restructured and now states only legacy moonshot-v1 rates) and Groq's five host rows (the rate table was withdrawn and the console pricing path 404s). Both are carried forward at their 2026-08-06 dates rather than re-dated, and both are in the coverage ledger.
- All other rows re-fetched and confirmed unchanged: OpenAI 5, Anthropic 7, Google 4, Mistral 3, Z.ai 4, Alibaba 3, xAI 2, Perplexity 2, MiniMax 1, Llama 2, plus the Together and Hugging Face host blocks.

2026-08-06

- Weekly refresh: all 41 model rows and all 3 host rows re-fetched against their sourced pages. Five rows moved on price, licence or context; the rest were confirmed unchanged.
- OpenAI cut GPT-5.6 Luna from $1.00/$0.10/$6.00 to $0.20/$0.02/$1.20 (–80%) and GPT-5.6 Terra from $2.50/$0.25/$15.00 to $2.00/$0.20/$12.00 (–20%), effective 2026-07-30. Sol is unchanged.
- Corrected last pass's long-context note on the GPT-5.6 rows. It said the long tier is '2x these rates'; the page's own numbers are 2x on input but 1.5x on output (Terra $4.00/$18.00, Luna $0.40/$1.80, Sol $10.00/$45.00). The rows now carry the actual figures, and say plainly that the page does not state the threshold that triggers the tier — the '>272K' we published last week was not on it.
- Filled cache-write rates for the GPT-5.6 rows ($6.25 / $2.50 / $0.25), which the pricing page does state and we had carried as null. GPT-5.5 and GPT-5.4 mini still show no cache-write rate on the page and stay null.
- Added Claude Mythos 5 — same $10/$50 pricing as Fable 5, same 1M window, limited availability through Project Glasswing.
- Added two sourced Anthropic notes on Fable 5's safety classifiers: a refusal returns HTTP 200 with `stop_reason: "refusal"` rather than an error, refused requests that produced no output are not billed, and retry is available server-side (`fallbacks`, beta), through SDK middleware, or by hand, with fallback credit covering the prompt-cache cost of switching. Mythos 5 is the same model without the classifiers, which is the substantive difference between the two rows.
- Mistral licences resolved against the models doc rather than left as 'check model card': Medium 3.5 is Modified MIT (v26.04), Small 4 Apache-2.0 (v26.03), Large 3 Apache-2.0 (v25.12). Devstral 2 stays 'partial' — the pricing page calls it Open but names no licence.
- Qwen rows re-sourced to Alibaba's own Model Studio pricing page instead of hosts. Both current rows are on limited-time promotions (Max 50% off, Plus 20% off), Plus is priced in two tiers with the step at 256K input, and no cached-input rate is stated first-party — the 90% cached discount we published last week is gone from the row.
- Anthropic cache minimum-token thresholds recorded per model (512 on Fable/Mythos/Opus 5, 1024 on Sonnet 5, Opus 4.8 and Sonnet 4.6, 4096 on Haiku 4.5), sourced to the prompt-caching doc.
- GLM-5.2 now carries a 1M window and 131K max output from the secondary comparison source, with the row saying explicitly that Z.ai's own pricing page still states no window.
- DeepSeek V4-Flash is now served as V4-Flash-0731 — retrained weights, same model ID, same $0.14/$0.28. No board value changed; the news entry flags it because a pinned ID moved underneath.
- Deferred Qwen3.8-Max (announced 2026-08-03): live per the announcement but absent from the Model Studio pricing page, and the promised open weights are not out. No rate, no licence, no row.
- Bulk-extraction use case rewritten — Luna's cut puts a frontier model within a rounding error of the cheap tier on input, which is the kind of change that quietly invalidates a routing rule.
- Review fix: the Qwen3.7-Max note contradicted itself, saying Together lists $0.13 cached and then that no cached rate is stated on any of the three pages. Both Qwen rows now say plainly what was withdrawn and why: the $0.25 and $0.35 cached rates came from hosts, and were dropped when the rows moved to Alibaba's own pricing rather than describe a lab's row with a host's cache rate.
- Review fix: Claude Mythos 5's window was sourced to a line about "Claude Mythos Preview" — which the launch doc names as the model Mythos 5 succeeds, so the citation was for a different model. Re-sourced to the launch doc, which states Fable 5 and Mythos 5 share specs, and the row now carries max_output 128000 and says the two are not the same model.

- Review fix: Mythos 5's best_for restated price and access instead of naming a use. It now says what the row is for — approved Project Glasswing security work — and its capability flags come from the launch doc rather than sitting as 'unverified'.
- Claude Fable 5 gains max_output 128000 from the same launch doc, and its watch_out now names the refusal behaviour a caller has to handle: HTTP 200 with stop_reason "refusal", not an error.

2026-07-30

- First board. 40 models across 12 labs (27 latest-generation, 13 previous), plus 3 inference hosts and hubs — every row carrying a source URL fetched on 2026-07-30.
- 12 of the 40 models ship open weights; MIT (DeepSeek V4, GLM-5 line) and Apache-2.0 (open Qwen lines) are the least restrictive, Llama and Kimi attach user-count and revenue thresholds.
- 18 of the 40 rows in that first board stated a context window on their sourced page. The rest are shown as unstated and omitted from the context chart rather than estimated — the chart states the count itself, computed.
- Added the caching cost panel: the same 50K-prompt job priced with and without caching, computed from the rate table rather than typed in.
- Added per-provider benefits and challenges, sourced to provider documentation.
- Published a coverage ledger naming what this pass could not source, rather than leaving gaps unexplained.
- Review fix: Moonshot, Alibaba, Meta and MiniMax had a host's pricing page (together.ai) recorded as their own provider pricing_url. A host's price is not the lab's price, so that misstated provenance. Moonshot and Alibaba now point at their own pages; Meta and MiniMax are null, because neither publishes a first-party rate we could reach — each says so in its notes. The validator now rejects a lab citing a host domain, which caught two more cases than the review found.
- MiniMax M3 corrected: the 1M window is priced in two tiers with the step at 512K input ($0.60/$2.40 above it), and the $0.30/$1.20 headline is a standing promotional 50% discount off that. Both now stated on the row.

Naderu

We train, release and run specialised models — and we are model-agnostic by policy. This board is how we show it.

BytesBrains Pte. Ltd.
naderu.com · contact@naderu.com