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 **2026-10-22**
- Refreshed **weekly**
- Next update **2026-10-29**
- **77** models · **15** labs · **3** hosts
- **263** source links, all fetched 2026-10-22
- 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.

CHAT ASSISTANT

Multi-turn conversation, the ordinary case

→ Claude Haiku 5.5 Anthropic

→ Gemini 3.8 Flash Google · Gemini

→ GPT-6 Luna OpenAI

The catch: the trap here is caching rather than headline price: all three cache, but each has a minimum below which a prompt is never cached and no error says so — 512 tokens on Claude Haiku 5.5, 1,024 on GPT-6 Luna, 4,096 on Gemini 3.8 Flash. Claude Haiku 5.5 replaces Haiku 4.5 on this card at $0.10/$0.50, GPT-6 Luna's exact price and a tenth of Haiku 4.5's, but only up to 100,000 input tokens — above that the whole request bills $0.50/$2.50 — and a long conversation crosses that line on its own. Two more things change on the swap: adaptive thinking is on by default, so a turn can open with thinking billed as output, and the same text counts as about 30% more tokens. Gemini 3.8 Flash's $0.75/$3.75 doubles to $1.50/$7.50 on 2027-01-01.

DEEP REASONING

Hard problems that reward thinking tokens

→ GPT-6 Astra OpenAI

→ Claude Fable 5.1 Anthropic

→ Kimi K3 Moonshot · Kimi

The catch: this is the one job where the saver posture is the wrong default, and the bill is shaped by how much the model thinks rather than how much you ask. Reasoning tokens bill as output — OpenAI's pricing page says so outright — so the output rate is the number that matters: $50.00 on GPT-6 Astra and Claude Fable 5.1, $15.00 on Kimi K3. The subtler difference is whether thinking can be turned down at all. Astra's effort starts at low, with no none; Fable 5.1's adaptive thinking is always on; Kimi K3 always reasons and defaults to max effort. None of the three lets you switch reasoning off, so a cheap test at low effort does not bound the production bill at the default — pin the effort you measured. Kimi K3's $15.00 is a third of the other two, on open weights, and flat across its 1M window.

SUMMARISATION

Long documents in, short faithful text out

- → Gemini 3.8 Flash `Google · Gemini`
- → DeepSeek-V4.1-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 2% at DeepSeek ($0.006 against $0.30). Second, the context window is a hard gate rather than a nice-to-have, and this is where GLM-4.7-FlashX's cheapest-on-the-board $0.07 input comes with a limit: Z.ai's model guide now states its window at 200K, a fifth of the other two picks. A corpus that fits is the cheapest summarisation on the board; one that does not has to be chunked first. DeepSeek-V4.1-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.

CONTENT WRITING

Drafting and rewriting to a house style

- → Claude Sonnet 5.5 `Anthropic`
- → GPT-6 Luna `OpenAI`
- → Mistral Small 4 `Mistral`

The catch: output is the product here — drafting and rewriting both return the full text, so this is the text job where the output rate hurts most, and a second 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 5.5, which replaces Sonnet 5 on this card at the same $2/$10, caches from 512 tokens where Sonnet 5 needed 1,024, so a short house-style prompt that paid full rate last week can now cache. Below the minimum there is no error to tell you. GPT-6 Luna at $0.10/$0.50 is the cheapest serious option here and carries no end date. Output is the product, so that $0.50 is the number doing the work.

EXTRACTION

Classify or pull fields out of a corpus

- → GLM-4.7-FlashX `Z.ai · GLM`
- → DeepSeek-V4.1-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. DeepSeek-V4.1-Flash bills $0.30/$1.20 at peak and halves off-peak — outside 01:00–04:00 and 06:00–10:00 UTC on weekdays, and all day on weekends and Chinese public holidays, which DeepSeek's page now counts as off-peak in full. The legacy deepseek-v4-flash name still answers from V4.1-Flash, so a pipeline pinned to it already changed model without changing a line. GLM-4.7-FlashX at $0.07/$0.40 and Mistral Small 4 at $0.15/$0.60 are the picks here with neither a clock nor a promotion attached. GPT-6 Luna at $0.10/$0.50 now sits between them on price — worth a run on your own schema before you assume the cheapest row is the right one. Claude Haiku 5.5, new on 2026-10-07 and pitched by Anthropic at exactly this work, matches Luna's $0.10/$0.50 up to 100K input tokens — with thinking on by default, so set effort low before you compare.

CODE COMPLETION

Fill-in-the-middle and next lines, at keystroke latency

- Codestral (v25.08) `Mistral`
- DeepSeek-V4.1-Flash `DeepSeek`

The catch: fill-in-the-middle is a distinct capability, not a smaller agentic-coding job, and only two rows on this board carry a first-party statement that they do it — which is why this card has two picks, not three. Mistral describes Codestral as built for "high-frequency completion, fill-in-the-middle"; DeepSeek's FIM endpoint is in beta and works in non-thinking mode only. The economics invert agentic coding: requests are short and there are a great many of them, so the input rate and latency decide it — Codestral is $0.30/$0.90, DeepSeek-V4.1-Flash $0.30/$1.20 at peak and half that off-peak. Keep the file context stable at the front of the prompt: Mistral's caching docs name FIM requests that reuse a long prefix as a case caching is for. And Codestral sits in Mistral's proprietary Premier tier despite the open-weight history of the name.

AGENTIC CODING

Long tool-use loops over whole repos

- Claude Sonnet 5.5 `Anthropic`
- GPT-6.1 Sol `OpenAI`
- Kimi K2.7 Code `Moonshot · Kimi`

The catch: output tokens dominate an agent loop's bill, and both lab picks on this card moved this week at unchanged list prices. GPT-6.1 Sol replaces GPT-6 Sol at $2.00/$10.00 with its cache read halved to $0.10 — the loop's re-sent context is exactly what that discounts — but its effort starts at low, so a loop that ran GPT-6 Sol at none now pays for thinking. Claude Sonnet 5.5 replaces Sonnet 5 at $2/$10, and it is not a drop-in swap for an existing loop: forced tool_choice returns 400 and thinking cannot be switched off, as on Opus 5.5. The output multiplier still varies more than people expect — 5x on Sonnet 5.5 and GPT-6.1 Sol, about 4x on Kimi K2.7 Code, 3x at xAI. Cache the stable prefix and cap max_tokens before you optimise anything else.

CODE REVIEW

Read a diff or a repository and say what is wrong

- Claude Opus 5.5 `Anthropic`
- GPT-6.1 Sol `OpenAI`
- DeepSeek-V4.1-Flash `DeepSeek`

The catch: review reads far more than it writes — a diff and the files around it in, a short list of findings out — so it shares summarization's arithmetic: the input rate and the cache are the bill, and the repository context is the part to cache. The rows differ on how. Claude Opus 5.5 caches from 512 tokens and its cache read is $0.20 — 0.05x its input, half the ratio of every earlier Opus; GPT-6.1 Sol, which replaces GPT-6 Sol here, caches from 1,024 tokens in implicit or explicit-only mode at $0.10, also 0.05x, entries living at least 30 minutes; DeepSeek-V4.1-Flash caches by default, best-effort, at $0.006 per 1M at peak. What does not carry over from summarization is the output side: a reviewer that reasons before it answers bills that thinking as output. Neither Opus 5.5 nor GPT-6.1 Sol can turn thinking off — Opus 5.5 defaults to medium effort, GPT-6.1 Sol's floor is low — so pin the effort you measured.

DOCUMENT UNDERSTANDING

OCR, scans, charts and forms

- Gemini 3.1 Flash-Lite `Google · Gemini`
- Qwen3.8-Max `Alibaba · Qwen`
- GPT-6 Luna `OpenAI`

The catch: image and audio tokens are counted differently by every provider — Gemini 3.1 Flash-Lite prices audio input at 2x text, Gemini 3.5 Flash-Lite at the text rate — 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. Gemini 3.1 Flash-Lite now has an earliest shutdown of 2027-05-07, with 3.5 Flash-Lite named as its replacement; GPT-6 Luna takes GPT-5.6 Luna's place here at $0.10 input.

SPEECH

Audio in, text out — the part of it this board prices

- Gemini 3.5 Flash-Lite `Google · Gemini`
- Gemini 3.1 Flash-Lite `Google · Gemini`

The catch: this board prices tokens, and speech is mostly sold by the minute, so the card is narrower than the job: it covers audio understanding on the two rows whose pricing page prices audio input explicitly. They split the way that matters. Gemini 3.5 Flash-Lite bills audio at its text rate, $0.30; Gemini 3.1 Flash-Lite is cheaper for text at $0.25 but charges $0.50 for audio. So for audio-heavy input 3.5 is the cheaper way in and the dearer way out — $2.50 output against $1.50 — and a long recording summarised briefly favours it while a transcript-length answer does not. Transcription and synthesis are separate products on a separate unit, and none is a row here: Mistral's Voxtral Mini Transcribe 2, for one, is $0.003 per minute of audio. Compare those per minute, not per token.

RAG AND GROUNDED ANSWERS

Answers built on passages you retrieved

- Claude Sonnet 5.5 `Anthropic`
- Gemini 3.8 Flash `Google · Gemini`
- GPT-6 Luna `OpenAI`

The catch: internal vector DB retrieval eliminates third-party search fees entirely, making prompt cache architecture the primary cost driver. The stable head (system prompt, schemas, extraction instructions) caches, while retrieved private chunks vary per query. Place stable instructions first and dynamic chunks second so prompt caching triggers: 512 tokens on Claude Sonnet 5.5 ($0.10 cache read, high needle-in-haystack adherence), 1,024 on GPT-6 Luna ($0.01 cache read, $0.10 base), and 4,096 on Gemini 3.8 Flash ($0.075 cache read, 1M context window). In enterprise RAG, strict source adherence and needle recall across 128K+ chunk contexts matter more than raw speed; models must synthesize evidence without hallucination or context dilution, and keep private company embeddings and text securely inside your boundary with zero external web leakage.

WEB SEARCH

Live internet retrieval, fresh facts and search grounding

- Sonar `Perplexity · Sonar`
- Gemini 3.8 Flash `Google · Gemini`
- Grok 4.7 `xAI · Grok`

The catch: live search grounding turns token pricing into a two-part bill: tokens for generation plus search fees per tool call or per 1,000 queries. Perplexity Sonar moved to the Agent API: $0.25 in / $2.50 out per 1M tokens plus $0.0025 per web search call ($2.50 per 1,000 searches) and $0.0005 per URL fetch — cheaper than the discontinued Chat Completions $5–$12 request fee, but an agent that executes multiple searches pays for each call. Gemini 3.8 Flash includes 5,000 free Google Search grounding queries per month across Gemini 3.x, then $14 per 1,000 requests on top of $0.75/$3.75 token rates. Grok 4.7 provides real-time X and web indexing at $2.00/$10.00 with tool execution overhead. Real-time web retrieval adds 500–1,500ms latency overhead and injects dynamic snippets that bust prompt prefix caching — unlike private RAG, search-grounded queries pay full base input rates on retrieved snippets.

COMPUTER USE

Vision and tools driving a real UI

- Claude Sonnet 5 `Anthropic`
- Claude Opus 5 `Anthropic`
- GPT-5.6 Terra `OpenAI`

The catch: the loop is what you pay for: every step sends a screenshot back in, so input — image input in particular — dominates, and the context grows with every turn. Anthropic's computer use follows standard tool-use pricing plus a fixed toolset-definition overhead on every request, and its docs describe pruning old screenshots on a fixed interval so the cached prefix stays identical between prunes — and, on Claude Fable 5.1, avoiding client-side pruning altogether, because removing an earlier screenshot invalidates every later thinking block. Claude Opus 5.5 and Claude Sonnet 5.5 are not on this card yet for a concrete reason: on the Claude API both refuse the earlier computer_20251124 tool and need computer_toolset_20260801, so a working loop has to move toolsets before it moves model. On OpenAI's side, computer use is listed as a tool on the three GPT-5.6 rows and GPT-6 Astra, so the choice there is price and effort rather than availability. Measure cost per completed task, not per call: a cheaper model that needs twice the steps is not cheaper.

CONTENT MODERATION

Classify against a policy

- → GPT-6 Luna `OpenAI`
- → Mistral Small 4 `Mistral`
- → GLM-4.7-FlashX `Z.ai · GLM`

The catch: check first whether you need a general model at all. OpenAI's moderation endpoint is free to use, and Mistral lists Mistral Moderation 2 as free, alongside Shieldstral 1.0, an Apache-2.0 moderation model you can host — none of them a row here, because none is a general model priced per token. OpenAI's endpoint does not classify audio. Reach for the picks on this card when the policy is yours rather than a vendor's category list: a general model reading your own rules is how you classify against them, and at moderation volume the input rate is the bill — GPT-6 Luna at $0.10, Mistral Small 4 at $0.15 with Apache-2.0 weights, GLM-4.7-FlashX at $0.07.

TYPED DECISIONS (JEV-CLASS)

State in, a calibrated choice out — no text generated

- → Perplexity Decider v1.2 `Perplexity · Sonar`
- → Perplexity Decider Flash v1.2 `Perplexity · Sonar`
- → Jev 1.14 `TypeSafe AI`
- → GPT-6 Luna · Decisions API `OpenAI`

The catch: this is a new kind of model, not a cheaper language model, and the price column misleads in both directions. Every row here bills input only because it generates nothing: you send state plus typed questions (yes or no, pick one of these, score on this rubric) and get back a probability per option. Per 1M input tokens: Perplexity's Decider v1.2 $0.02 and Flash v1.2 $0.01, Jev 1.14 $0.042, decision-machine-1 and meraGPT's Decider 1 $0.04, and OpenAI's Decisions API on GPT-6 Luna $0.10 — a ten-fold spread for the same shape of answer. So the fair comparison is per decision, against the language-model call you make today plus the parsing and retries around it. Three things decide it. The per-request budget and inputs: text, JSON and images on Perplexity's v1.2 (its input budget is unstated on the pages fetched this pass), text and images on OpenAI, 64K text on Jev 1.14 (262K with the opt-in long-context header), 4,096 tokens on Decider 1. Whether the probabilities are calibrated on your data rather than the vendor's — every speed multiple on these pages, OpenAI's “10x faster” included, is the vendor's own measurement. And maturity: OpenAI's is a beta with one model, Jev is in early access, and Alibaba's decision-model-preview is still free with no list rate, so it is not a row. A job that needs prose, facts it would have to look up, or arithmetic still belongs to a language model.

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, Grok and OpenAI it is not even the advertised price — Gemini and Grok double per token above 200K input, OpenAI's GPT-5.5, 5.6 and 6 rows bill 2x input and 1.5x output for the whole request above 272K, and xAI states the higher rate applies to every token, 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 window is now stated by Meta's own model card — but no public host on this board serves it for inference: Together lists it only for fine-tuning, and Groq dropped it from its free and developer tiers.

OPEN WEIGHTS

Weights you can download, run and fine-tune

"open weights" is not one licence, and this pass made that easier to read rather than harder. MIT with no strings: DeepSeek, GLM-5, GLM-5.2 and GLM-5.3-Flash, each stated on its own Hugging Face card. Apache-2.0: the open Qwen lines and Meta's Muse Glimmer 30B. Then the licences of their own name, each with a clause worth reading: GLM-5.3's adds a Z.ai security review for model-as-a-service businesses above $10B revenue; Kimi K3's requires a separate agreement above $20M MaaS revenue; MiniMax M3's Community License is non-commercial by default and needs written authorisation above $20M yearly revenue; Qwen3.8-Max's names a custom licence the card does not print; Llama attaches a 700M-MAU threshold. 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 Llama rows are now enterprise-only, so the fastest published figures a self-serve customer can buy are GPT OSS 20B and 120B.

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 they do the price differs by model more than by provider: 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, while Gemini 3.5 Flash-Lite bills audio at its $0.30 text rate. Budget audio as its own line, and read the row rather than assuming the family's rule.

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 a twelfth of the input rate precisely because what you send and the responses you get back are used to train future Meta models — 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 usually a flat 50% on both input and output — OpenAI, Anthropic, Google and Mistral all state that rate. Two labs price it differently: xAI gives Grok 4.3 20% (and supports no batch at all on Grok 4.7 or 4.6), and Moonshot bills K2.7 Code and K2.6 at 60% of standard, a 40% discount, with no batch for K3. 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: Anthropic's batch discount does not stack with fast mode, and it does not apply to Managed Agents sessions at all.

WEB SEARCH NATIVE

Direct internet queries and provider search grounding

identifies models that execute live web queries directly or through integrated provider grounding tools (Perplexity Sonar, Gemini with Google Search Grounding, Grok live search). Hosted search is billed per tool call ($0.0025 on Sonar) or per 1,000 queries ($14 past free tier on Gemini) on top of token generation, adds 500–1,500ms network latency, and breaks prompt prefix caching because retrieved snippets differ on every call.

LONG-CONTEXT RETRIEVAL

High-fidelity needle recall across 128K+ token windows

identifies models that maintain >= 99% retrieval accuracy across >= 128k token context windows without context dilution or needle loss. Context window size is not retrieval capability: a 1M-token window does not guarantee accurate synthesis over large vector chunk sets unless the model maintains source adherence and low hallucination across its full context without middle-loss degradation.

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

$0.10/$0.50 is now a price point, not a row: GPT-6 Luna set it for a frontier lab, and Claude Haiku 5.5 matched it on 2026-10-07 — up to 100K input tokens, five times that above. DeepSeek-V4.1-Flash at $0.15/$0.60 still wins on cache: $0.003 off-peak and $0.006 at peak, against $0.01 on both. GLM-4.7-FlashX's $0.07 input is still the cheapest first-party language-model row, with a 200K window. If the output is a label rather than text, the typed-decisions card is the cheaper question to ask: $0.02 to $0.10 per 1M input with no output charge. Grok 4.3 and Kimi K3 are unchanged.

BALANCED

The cheapest row that clears your bar

the default posture, and the one the job cards assume. The $2/$10 price point holds four rows, and since 2026-10-07 the two current ones match on cache as well: Claude Sonnet 5.5's cache read fell from $0.20 to $0.10, GPT-6.1 Sol's level, eight days after OpenAI set it. Sonnet 5.5 caches from 512 tokens to GPT-6.1 Sol's 1,024; GPT-6.1 Sol's effort starts at low, and Sonnet 5.5's thinking cannot be switched off. Gemini 3.8 Flash at $0.75/$3.75 is still here and still doubles on 2027-01-01. GPT-5.6 Sol's $4.00/$20.00 is still labelled promotional through at least 2026-11-21.

QUALITY

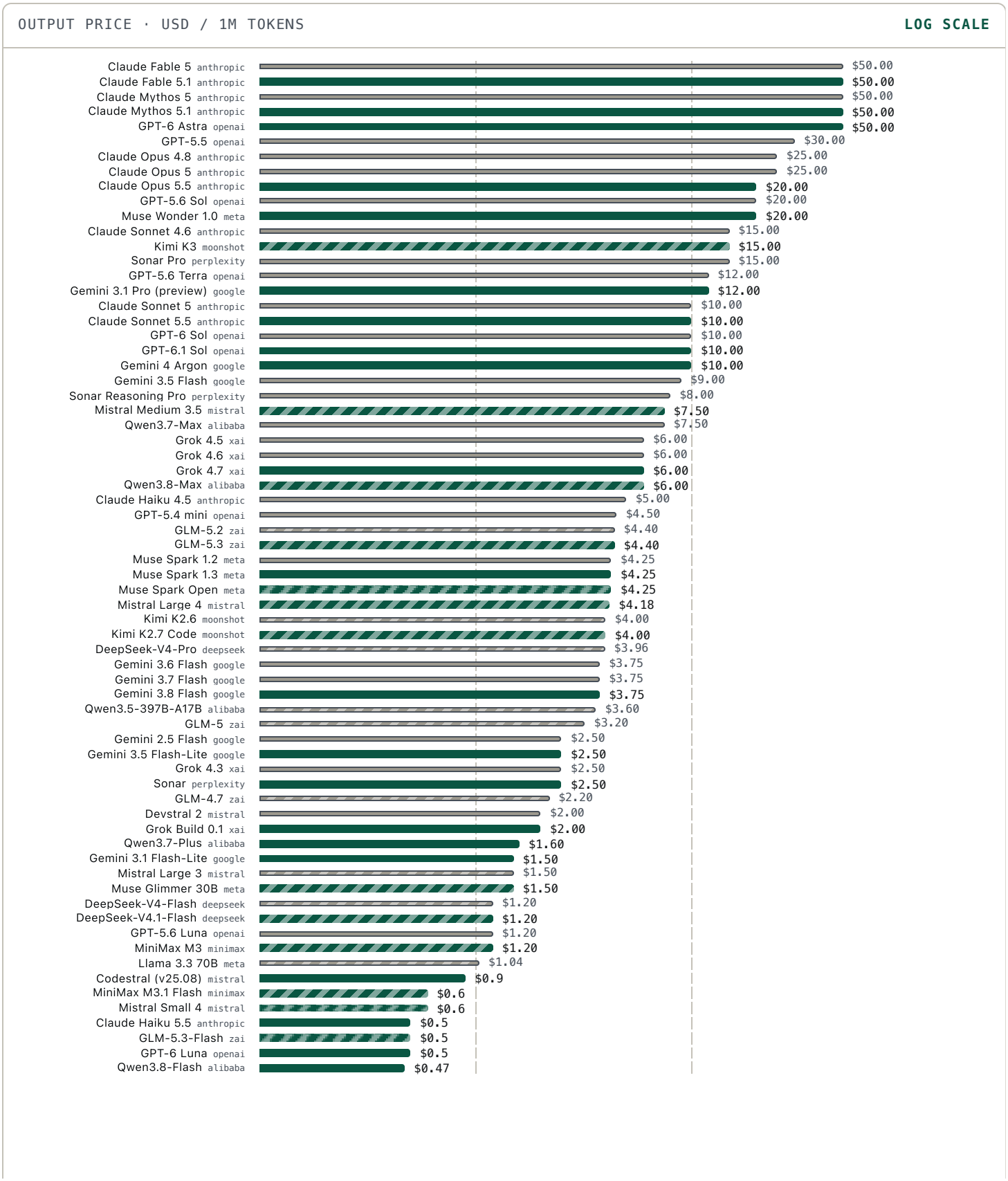
Top of the capability range, bill accepted

GPT-6 Astra and Claude Fable 5.1 and Mythos 5.1 share the top of the board at $10.00/$50.00, then Claude Opus 5.5 — new this pass, and now Anthropic's recommended starting point — at $4.00/$20.00, a fifth below Opus 5's $5.00/$25.00. The cache read separates the top rows: $0.25 on the Claude 5.1 rows, $0.20 on Opus 5.5, $1.00 on Astra. Three things to settle before committing: Mythos 5.1 is allocated rather than sold, so a roadmap cannot depend on it; the Fable line and now Opus 5.5 ship safety classifiers that decline with HTTP 200 and stop_reason "refusal", which is an integration change and not a footnote; and Opus 5.5's thinking cannot be disabled.

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.

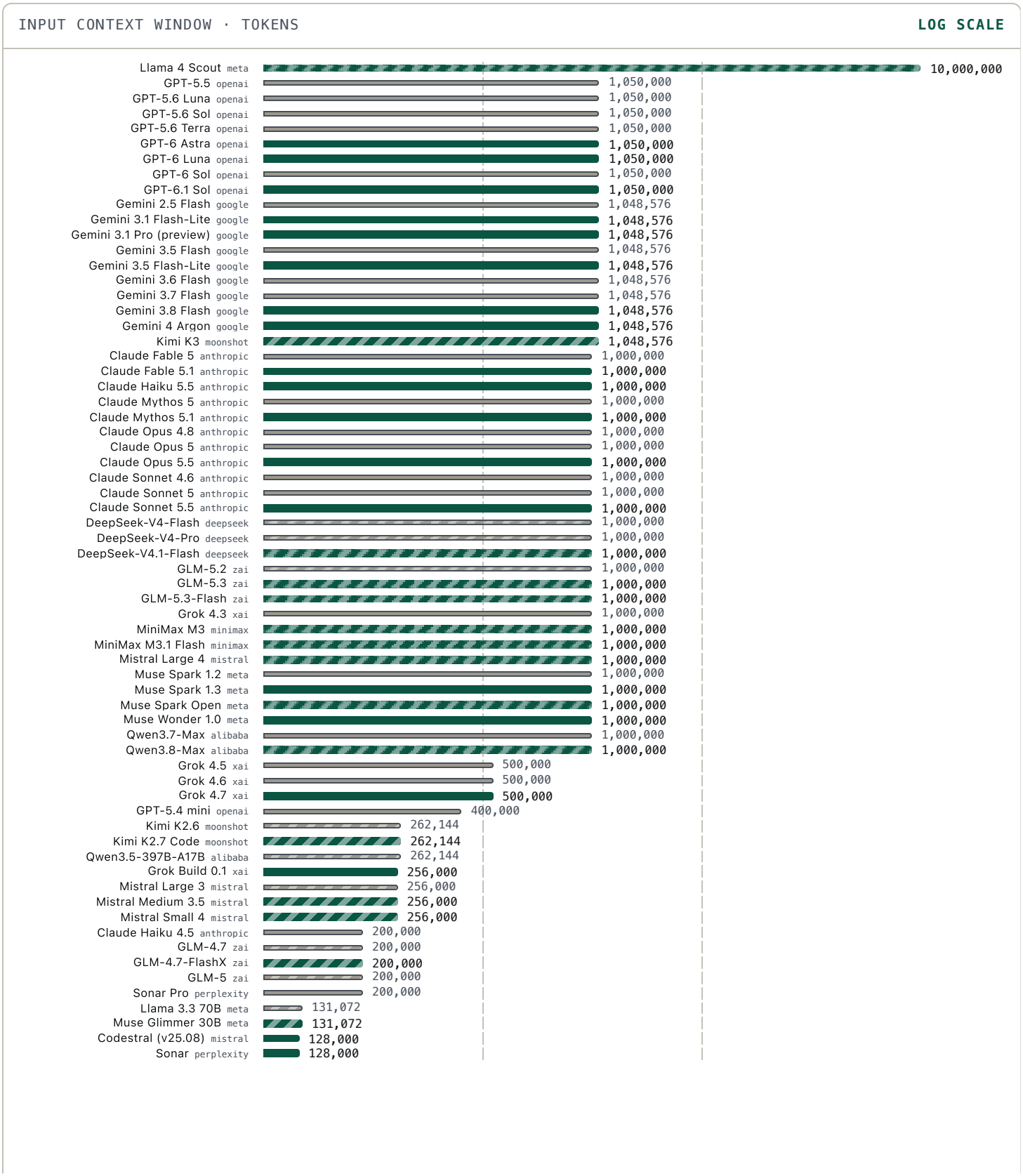


GLM-4.7-FlashX zai
$0.4
7 typed-decision models are left off this chart: they generate no text, so they bill input only and have no output to compare. See the typed-decisions card.
USD / 1M OUT
$0.10
$1
$10
$100
Latest generation
Previous generation
Open weights — self-hostable, fine-tunable

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.



3 of 70 language models are absent from this chart because their sourced pages do not state a context window. We show nothing rather than a guess — see the coverage ledger. The 7 typed-decision models are left off too: what they state is a per-request budget for a different kind of API — see the typed-decisions card.

TOKENS IN 100K 1M 10M

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.

77 of 77 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**	—	262K	—	not stated	?	?	Yes	unverified	Open	Apache-2.0 — stated on the Hugging Face model card	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 · native window on the Hugging Face card, "extensible up to 1,010,000"	src1 src2 2026-10-22

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
Qwen3.7-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, and Qwen3.8-Max is cheaper on both sides ($2.00/$6.00). Both hosts are cheaper still — shop it	Model Studio Singapore rate. The limited-time 50%-off label this row carried is gone from the pricing page, and $2.50/$7.50 now reads as the standard price. 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, $1.50/$0.30 cached/$4.50 on Together — up from $1.25/$0.13/$3.75 last pass — and that $0.30 is Together's rate for Together's cache, not Alibaba's · 65,536 max output carried from an earlier pass: no Alibaba page restated it this pass, and OpenRouter's listing says up to 131,072 completion tokens	src1 src2 src3 src4 2026-10-22
Qwen3.7-Plus ALIBABA · QWEN · LATEST	$0.4	—	**$1.60**	—	—	—	not stated	?	?	Yes	both	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 / $4.80 out above 256K. Together lists it at $0.32/$1.28, below Alibaba's own rate · window not stated; input is priced in two tiers with the step at 256K	src1 src2 2026-10-22

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-Flash ALIBABA · QWEN · LATEST	$0.15	—	**$0.47**	—	—	—	not stated	?	?	?	both	Not verified		Alibaba's cheapest current-generation Qwen tier	Priced, and little else: Alibaba states no window and no capabilities for it, only that it can cache. Treat it as unmeasured until you have run it	Model Studio international rate, one tier up to 1M input tokens. The cached-hit rate is excepted from Alibaba's 10%/20% multipliers and shown only in the console. Together now lists it at $0.09/$0.28, below the first-party rate · not stated — the price tier runs to 1M input tokens, which is a billing band, not a window	src1 src2 2026-10-22
Qwen3.8-Max ALIBABA · QWEN · LATEST	$2.00	$0.25	**$6.00**	—	1M	131K	effort knob	Yes	?	Yes	both	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 cached column is the implicit rate; the explicit read is $0.17 · 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 src3 2026-10-22
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 on 2026-09-01 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 2026-10-22

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 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 Claude Opus 5.5 pays 0.05x and every other Claude row 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 src4 2026-10-22
Claude Haiku 4.5 ANTHROPIC · PREVIOUS	$1.00	$0.1	**$5.00**	−50%	200K	64K	effort knob	Yes	—	Yes	both	Closed		Fast interactive turns that still need real tool use	Legacy since Claude Haiku 5.5 shipped on 2026-10-07, at ten times Haiku 5.5's rate on prompts up to 100K tokens. Retirement is committed no sooner than 2026-10-15 and no notice has been given; Anthropic promises at least 60 days, so it cannot retire on that date. 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, stated on its own model page, which now labels it Legacy	src1 src2 src3 src4 2026-10-22
Claude Haiku 5.5 ANTHROPIC · LATEST	$0.1	$0.01	**$0.5**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Anthropic's high-volume tier — classification, routing, extraction and subagent work — at a tenth of Haiku 4.5's price below 100K tokens	Not a drop-in swap for Haiku 4.5 code: adaptive thinking is on by default at medium effort, so responses can begin with thinking blocks billed at the output rate; budget_tokens, assistant prefill and any non-default temperature, top_p or top_k return 400. The same text counts as about 30% more tokens. Priority Tier is not supported	for prompts up to 100,000 tokens; above that the whole request bills $0.50 in / $0.05 cache read / $0.625 5m write / $2.50 out — five times the short-prompt rate. 1h cache write $0.20 ($1.00 above 100K). Released 2026-10-07 · 300K max output on the Message Batches API with a beta header	src1 src2 src3 src4 src5 src6 src7 2026-10-22

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 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 on 2026-09-01 and still invite-only. 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. Retirement not sooner than 2027-06-09	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 2026-10-22
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 Claude Opus 5.5 pays 0.05x and every other Claude row 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 src4 2026-10-22
Claude Opus 4.8 ANTHROPIC · PREVIOUS	$5.00	$0.5	**$25.00**	−50%	1M	128K	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 src3 2026-10-22
Claude Opus 5 ANTHROPIC · PREVIOUS	$5.00	$0.5	**$25.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Long agent runs already validated on it, and code that relies on disabling thinking or forcing a tool	Now Legacy on Anthropic's models page: Claude Opus 5.5 is $4/$20 against this row's $5/$25, with a cache read at half the rate. Retirement not sooner than 2027-07-24, so there is no clock — but Opus 5.5 breaks code that disables thinking or forces tool_choice, so migrate on a test, not a find-and-replace	1h cache write $10.00; fast mode $10/$50 · full 1M window at standard rates; 128K max output per request	src1 src2 src3 src4 2026-10-22

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 Opus 5.5 ANTHROPIC · LATEST	$4.00	$0.2	**$20.00**	−50%	1M	128K	always on	Yes	—	Yes	both	Closed		Anthropic's recommended starting point for most work — long agentic coding and knowledge work, a fifth cheaper than Opus 5 on every column	Not a drop-in swap for Opus 5. Thinking cannot be disabled (400), forced tool_choice any/tool returns 400, the default effort drops from high to medium, and text between tool calls now arrives in thinking blocks. It also ships safety classifiers that decline with HTTP 200 and stop_reason "refusal", as the Fable line does	1h cache write $8.00; fast mode $8/$40. Cache reads bill at 0.05x input — half the 0.1x Opus 5 and every earlier Opus pays · full 1M window at standard rates; up to 300K output on the Batch API with a beta header	src1 src2 src3 src4 src5 2026-10-22
Claude Sonnet 4.6 ANTHROPIC · PREVIOUS	$3.00	$0.3	**$15.00**	−50%	1M	128K	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 src3 2026-10-22
Claude Sonnet 5 ANTHROPIC · PREVIOUS	$2.00	$0.2	**$10.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Existing Sonnet 5 integrations that rely on forced tool use or on switching thinking off — both of which Sonnet 5.5 refuses	Legacy on Anthropic's overview since Sonnet 5.5 shipped on 2026-09-28 at the same list price. Not deprecated — retirement no sooner than 2027-06-30 — but since 2026-10-07 its $0.20 cache read is twice Sonnet 5.5's $0.10, and it caches from 1,024 tokens where Sonnet 5.5 caches from 512	$2/$10 is the standard rate; the rise to $3/$15 once announced for 2026-09-01 was withdrawn. 1h cache write $4.00 · full 1M window at standard rates; 128K max output per request	src1 src2 src3 src4 2026-10-22

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 5.5 ANTHROPIC · LATEST	$2.00	$0.1	**$10.00**	−50%	1M	128K	effort knob	Yes	—	Yes	both	Closed		Anthropic's balance of speed and capability, at Sonnet 5's price with half its cache read	Not a drop-in swap for Sonnet 5 code: forced tool use (tool_choice any or tool) returns 400; thinking cannot be switched off — the lowest setting, between_tools, only drops the up-front pass; text between tool calls now arrives in thinking blocks; and on the Claude API the earlier computer_20251124 tool is refused	released 2026-09-28 at Sonnet 5's price; cache read cut from $0.20 to $0.10 on 2026-10-07. 1h cache write $4.00. The pricing page's model table still prints $0.20 in the cache-hit column, while the same page's multiplier section and the 2026-10-07 release note both state $0.10 — this row follows the dated change · full 1M window at standard rates; 128K max output per request, 300K on the Batch API with a beta header	src1 src2 src3 src4 src5 2026-10-22
DeepSeek-V4-Flash DEEPSEEK · DEPRECATED	$0.3	$0.006	**$1.20**	—	1M	384K	effort knob	—	—	Yes	automatic	Open	MIT — stated on the Hugging Face model card	Nothing new — retired, and listed here so the retirement is on the record	Retired on 2026-09-10. The name still answers, from DeepSeek-V4.1-Flash — a different model behind the same string. Move to deepseek-flash and re-run your evals	retired: requests are served by DeepSeek-V4.1-Flash-1017 and billed at the Flash price, stated on DeepSeek's own quick-start page. Kept as a row so a config still pinning the old name prices correctly	src1 src2 src3 src4 2026-10-22
DeepSeek-V4-Pro DEEPSEEK · PREVIOUS	$1.32	$0.044	**$3.96**	—	1M	384K	effort knob	—	—	Yes	automatic	Open	MIT — stated on the Hugging Face model card	Frontier-adjacent capability at a fraction of frontier price, with open weights — still a live endpoint after the walked-back cutover	Was marked for retirement on 2026-09-14; DeepSeek's changelog now says the API continues after that date with billing unchanged, until further notice, and the pricing page still lists it at its own rate. The original release note still describes the cutover. No V4.1-Pro has shipped	peak rate; off-peak is half, $0.66/$0.022/$1.98, outside 01:00–04:00 and 06:00–10:00 UTC, Monday to Friday, excluding Chinese public holidays. Still billed at this model's own rate — DeepSeek walked back the 2026-09-14 plan to route the name to V4.1-Flash	src1 src2 src3 src4 src5 src6 2026-10-22

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-V4.1-Flash DEEPSEEK · LATEST	$0.3	$0.006	**$1.20**	—	1M	384K	effort knob	Yes	?	Yes	automatic	Open	MIT — stated on the Hugging Face model card	A 1M window with vision, tools and fill-in-the-middle at $0.30/$1.20, and the cheapest cache read on the board	Released 2026-09-10. V4-Flash is retired and routes here; V4-Pro does not — DeepSeek kept serving it at its own rate after walking back the 2026-09-14 cutover. Pin deepseek-flash rather than a legacy flash name. DeepSeek says V4.1-Flash outperforms V4-Pro; that is their claim, not a result on this board	peak rate. Off-peak is half — $0.15/$0.003/$0.60 — every hour outside 01:00–04:00 and 06:00–10:00 UTC, Monday to Friday, excluding Chinese public holidays. Since 2026-10-17 deepseek-flash is served by the DeepSeek-V4.1-Flash-1017 checkpoint with no calling change; the legacy names deepseek-v4-flash and deepseek-v4-flash-vision-exp are retired and answer from 1017 at this price. The price table itself renders client-side and was not re-read server-side this pass — three secondary reports agree it is unchanged, and Together lists 1017 at $0.30/$0.006/$1.20	src1 src2 src3 src4 src5 2026-10-22
Gemini 2.5 Flash GOOGLE · GEMINI · PREVIOUS	$0.3	$0.03	**$2.50**	−50%	1.05M	65K	effort knob	Yes	Yes	Yes	automatic	Closed		Existing projects that already call it — Google no longer lets new projects in	Since 2026-09-18 Google limits the 2.5 models to users who have actively used them before; they are not deprecated and keep serving "until further notice", but a new project may not get access at all. Google points new work at 3.5 Flash-Lite or 3.8 Flash. Implicit caching is best-effort	audio input $1.00 · stated on the model's own page, which gives the exact figure rather than "1M"	src1 src2 src3 src4 src5 2026-10-22
Gemini 3.1 Flash-Lite GOOGLE · GEMINI · LATEST	$0.25	$0.025	**$1.50**	−50%	1.05M	65K	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. Google's deprecations page now gives it an earliest shutdown of 2027-05-07, with Gemini 3.5 Flash-Lite as the replacement	audio input $0.50 in / $0.05 cached · stated on the model's own page, which gives the exact figure rather than "1M"	src1 src2 src3 2026-10-22

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 3.1 Pro (preview) GOOGLE · GEMINI · LATEST	$2.00	$0.2	**$12.00**	−50%	1.05M	65K	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	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 · stated on the model's own page, which gives the exact figure rather than "1M"	src1 src2 src3 2026-10-22
Gemini 3.5 Flash GOOGLE · GEMINI · PREVIOUS	$1.50	$0.15	**$9.00**	−50%	1.05M	65K	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	stated on the model's own page, which gives the exact figure rather than "1M"	src1 src2 src3 2026-10-15
Gemini 3.5 Flash-Lite GOOGLE · GEMINI · LATEST	$0.3	$0.03	**$2.50**	−50%	1.05M	65K	effort knob	Yes	Yes	Yes	both	Closed		Audio-heavy input, where it bills audio at the text rate	Dearer than Gemini 3.1 Flash-Lite on text in ($0.30 against $0.25) and on everything out ($2.50 against $1.50), and cheaper only on audio in. Google calls it the most cost-effective 3.5 model, which is a claim about the 3.5 line, not about the board	one input rate for text, image, video and audio — no audio premium, where Gemini 3.1 Flash-Lite charges $0.50 for audio against $0.25 for text. Cache storage $1.00 per 1M/hr · stated on the model's own page, which gives the exact figure rather than "1M"	src1 src2 src3 src4 2026-10-22
Gemini 3.6 Flash GOOGLE · GEMINI · PREVIOUS	$0.75	$0.075	**$3.75**	−50%	1.05M	65K	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 · stated on the model's own page, which gives the exact figure rather than "1M"	src1 src2 src3 2026-10-22
Gemini 3.7 Flash GOOGLE · GEMINI · PREVIOUS	$0.75	$0.075	**$3.75**	−50%	1.05M	65K	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 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 · 1,048,576 in / 65,536 out, stated on the model's own page — the latest-model page rounds them to 1M / 64K	src1 src2 src3 src4 2026-10-15

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 3.8 Flash GOOGLE · GEMINI · LATEST	$0.75	$0.075	**$3.75**	−50%	1.05M	65K	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 · 1,048,576 in / 65,536 out, stated on the model's own page — the latest-model page rounds them to 1M / 64K	src1 src2 src3 src4 src5 2026-10-22
Gemini 4 Argon GOOGLE · GEMINI · LATEST	$2.00	$0.1	**$10.00**	−50%	1.05M	65K	always on	?	?	Yes	automatic	Closed		Long-horizon software engineering and multi-step agents inside Google's stack, at $2/$10 until the 1 December doubling	The introductory rate has an expiry date: budget $4/$20 from 2026-12-01, and the cached rate doubles with it ($0.10 to $0.20). Paid tier only, and every figure above is Google's own — no third party has replicated the launch benchmarks	introductory through 2026-11-30; $4.00/$0.20/$20.00 from 2026-12-01. Not available on the free tier. No length tiering: one rate across the whole window. Grounding with Google Search bills apart: 5,000 free search requests a month, then $14 per 1,000 · 1,048,576 unified input-plus-output tokens; default synchronous max output is 65,536, rising to the full 1,048,576 with long decode, both stated on the model's own page	src1 src2 2026-10-22
Decider 1 MERAGPT · LATEST	$0.04	–	**free**	–	4K	–	—	—	—	—	unverified	Closed		A System One-schema endpoint for short states and small label sets	A 4,096-token cap and at most ten options per choice question — a sixteenth of Jev's per-request budget. New on 2026-09-22	billed per input token only, output $0.00. Up from $0.03 last pass · 4,096 tokens per request, state and questions together	src1 2026-10-22
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). Groq lists Llama 3.3 70B Versatile on console.groq.com/docs/models, now tagged Enterprise, with "Contact Sales" where a rate would go — so Groq publishes no price for it · max output from Groq's models page (2026-08-06); the window is the model's	src1 src2 2026-10-22

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 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 neither Together nor Groq's public tiers serve it for inference. The 10M window is now stated by Meta's own model card, after five passes on a secondary source that did not state it	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; it still appears there only under fine-tuning, now at $2.10 per 1M tokens for supervised fine-tuning with a $6.00 minimum. Groq deprecated it for free and developer tiers on 2026-07-17. Fewer hosts list this model than the 10M-window headline suggests · 10M stated by Meta's own Llama 4 model card	src1 src2 2026-10-15
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 2026-10-22
Muse Spark 1.2 META · LLAMA & MUSE · PREVIOUS	$1.25	$0.15	**$4.25**	—	1M	—	always on	Yes	?	Yes	automatic	Closed		Coding agents on a first-party Meta endpoint, flat-priced across 1M tokens	Superseded by Muse Spark 1.3 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 permission to use the prompts, code, files and context you send, plus the responses you get back, to train future Meta models, at 150 RPM against the standard tier's 3,000. Web-search grounding is $2.50 per 1,000 queries on top · the models page states 1,048,576; the board keeps its rounded 1M	src1 src2 src3 2026-10-22

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 1.3 META · LLAMA & MUSE · LATEST	$1.25	$0.15	**$4.25**	—	1M	—	always on	Yes	?	Yes	automatic	Closed		Coding agents on a first-party Meta endpoint, flat-priced across 1M tokens	The contributor tier is a twelfth of the input rate and a twenty-first of the output rate because it is a different trade: the prompts, code, files and context you send, plus the responses you get back, train future Meta models, at 150 RPM against the standard tier's 3,000.	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 permission to use the prompts, code, files and context you send, plus the responses you get back, to train future Meta models · the models page states 1,048,576; the board keeps its rounded 1M	src1 src2 src3 2026-10-22
Muse Spark Open META · LLAMA & MUSE · LATEST	$1.25	$0.15	**$4.25**	—	1M	—	not stated	Yes	Yes	Yes	unverified	Open		Self-hosted or host-routed Meta frontier coding work where the closed API's terms do not fit	Open weights are not open terms: read the licence file beside the weights before building, and price the host rather than assuming Meta's API rate card applies	host price (Together AI) — Meta serves no first-party endpoint for the open release; the weights (816B MoE, 42B active) are downloadable. Same rate as closed Spark on Meta's API, so the choice is custody and licence, not price · 1M stated by the host; Meta's overview does not restate it. Max output not stated	src1 src2 src3 2026-10-22
Muse Wonder 1.0 META · LLAMA & MUSE · LATEST	$4.00	$0.2	**$20.00**	—	1M	—	effort knob	Yes	Yes	Yes	unverified	Closed		Meta's most capable model for complex software engineering and long-horizon agents	In preview, and Meta's second closed model — the Llama licence does not apply. Muse Wonder Full, the higher-capability variant, is Muse Code subscription only, so max reasoning effort on the API model does not select it	standard mode only — there is no contributor tier for Wonder. No long-context premium, like Spark · 1,048,576 stated on the models page; the board keeps its rounded 1M. Max output not stated	src1 src2 2026-10-22
decision-machine-1 MILLISECONDS.AI · LATEST	$0.04	—	**free**	—	—	—	—	Yes	—	—	unverified	Closed		Typed decisions with a free evaluation allowance large enough to test a real workload	Spend-gated production limits, and terms that restrict using it to benchmark a competing model. System One-compatible route live since 2026-09-21	billed per input token only, output free. 125M free input tokens a month on test keys; production draws on purchased credits from $10. Images are priced per 1,000 on its native routes · text capped at 20,000 characters per request rather than a token window	src1 src2 2026-10-22

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
MiniMax M3 MINIMAX · LATEST	$0.3	$0.06	**$1.20**	—	1M	262K	not stated	Yes	?	Yes	automatic	Open	MiniMax Community License — non-commercial grant; commercial use needs attribution, and written authorisation above $20M yearly revenue	Cheap general-purpose model with an 80% cached-input discount	"Permanent" is MiniMax's word for the 50% discount, not a contract — budget the $0.60/$2.40 it is struck through from. Above 512K input the rate doubles. The weights are open under the MiniMax Community License, which is non-commercial by default: commercial use needs "Built with MiniMax M3" attribution and, above $20M yearly revenue, written authorisation	labelled "Permanent 50% off" on MiniMax's own pricing page, from $0.60/$2.40. Above 512K input: $0.60 in / $0.12 cached / $2.40 out. Priority tier 1.5x · 1M stated on MiniMax's own text-generation doc and priced in two tiers with the step at 512K input. The 262,144 max output is from OpenRouter; MiniMax states none	src1 src2 src3 src4 src5 src6 2026-10-22
MiniMax M3.1 Flash MINIMAX · LATEST	$0.15	$0.015	**$0.6**	—	1M	—	not stated	?	?	?	automatic	Open	minimax-community, as named on the Hugging Face model card — the licence file was not re-read this pass	MiniMax's cheapest row — half of M3 with a $0.015 cache read, and open weights to self-host	The weights landed on Hugging Face on 2026-10-23 under minimax-community. M3's MiniMax Community License is non-commercial by default, so read this model's own licence file before self-hosting. Capability columns are unstated rather than absent; the pricing page names none	listed on MiniMax's own pricing page at half of M3's rate in and out ($0.30/$1.20), and a quarter of its cache read. No priority tier is listed for it, unlike M3. The 2026-10-19 GA date is from secondary coverage; the pricing page carries no date · up to 1,000,000 tokens, stated on the Hugging Face model card; MiniMax's pricing page states no window or max output	src1 src2 src3 2026-10-22

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%	128K	–	—	—	?	Yes	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 on 2026-09-03. It is the −90% discount the other two pages describe only in prose · window stated on the model's own page; max output not stated	src1 src2 src3 src4 src5 2026-10-22
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 2026-07-31. 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 2026-10-15

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%	256K	—	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 on 2026-09-03. It is the −90% discount the other two pages describe only in prose · window stated on the model's own page; max output not stated	src1 src2 src3 src4 src5 2026-10-22
Mistral Large 4 MISTRAL · LATEST	$1.36	$0.14	**$4.18**	—	1M	—	not stated	Yes	?	Yes	unverified	Open		EU-hosted frontier work with function calling, agents and structured output	Public preview since 2026-10-06: 52B active of 1.05T parameters, docs badge Open but the weights ship end-October with a licence file nobody has read yet — this row prices the API, not self-hosting. Licence unknown until the weights land	standard rate on Mistral's own pricing page. Third-party coverage reports the public preview billing half that; the first-party page states no preview rate, so this row prints the stated one · 1M window stated on the model's own page; max output not stated	src1 src2 src3 2026-10-22
Mistral Medium 3.5 MISTRAL · LATEST	$1.50	$0.15	**$7.50**	−50%	256K	—	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 on 2026-09-03. It is the −90% discount the other two pages describe only in prose · window stated on the model's own page; max output not stated	src1 src2 src3 src4 src5 2026-10-22

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 Small 4 MISTRAL · LATEST	$0.15	$0.015	**$0.6**	−50%	256K	—	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 on 2026-09-03. It is the −90% discount the other two pages describe only in prose · window stated on the model's own page; max output not stated	src1 src2 src3 src4 src5 2026-10-22
Kimi K2.6 MOONSHOT · KIMI · PREVIOUS	$0.95	$0.16	**$4.00**	−50%	262K	—	effort knob	Yes	?	Yes	automatic	Open	Modified MIT — stated in the Hugging Face LICENSE	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 src2 2026-10-15
Kimi K2.7 Code MOONSHOT · KIMI · LATEST	$0.95	$0.19	**$4.00**	−50%	262K	—	always on	Yes	?	Yes	automatic	Open	Modified MIT — stated in the Hugging Face LICENSE	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 src2 src3 2026-10-15
Kimi K3 MOONSHOT · KIMI · LATEST	$3.00	$0.3	**$15.00**	—	1.05M	—	always on	Yes	?	Yes	automatic	Open	Kimi K3 License — a separate agreement above $20M MaaS revenue over 12 months; on-screen credit above 100M MAU or $20M monthly revenue	A 1M window with no long-context premium, and weights you can host	The weights ship under the Kimi K3 License, not the Modified MIT the K2 line carries: a model-as-a-service operator above $20M revenue over any 12 months needs a separate agreement with Moonshot. Reasoning is always on and defaults to max effort. Batch is not offered on K3	flat across the full window — no length tiering. cache write $3.00 at the default 5-minute TTL, $6.00 for 1 hour; a hit refreshes the entry at no write charge	src1 src2 src3 src4 2026-10-22
GPT-5.4 mini OPENAI · PREVIOUS	$0.75	$0.075	**$4.50**	−50%	400K	128K	effort knob	Yes	?	Yes	automatic	Closed		Cheap tier if you are already pinned to it	GPT-6 Luna is $0.10/$0.50 against this row's $0.75/$4.50, with a window 2.6x larger. Kept for code that pinned it	272,000 max input tokens, stated on the model's own page	src1 src2 2026-10-22

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-5.5 OPENAI · PREVIOUS	$5.00	$0.5	**$30.00**	−50%	1.05M	128K	effort knob	Yes	?	Yes	automatic	Closed			Dearer than every GPT-5.6 and GPT-6 row except Astra on output ($30.00), with an older caching scheme. Nothing on OpenAI's pages recommends it; it remains for code that pinned it	above 272K input tokens the whole session bills at 2x input and 1.5x output: $10.00/$1.00/$45.00 · stated on the model's own page	src1 src2 2026-10-22
GPT-5.6 Luna OPENAI · PREVIOUS	$0.2	$0.02	**$1.20**	−50%	1.05M	128K	effort knob	Yes	Yes	Yes	both	Closed		High-volume work that still needs vision and tools	GPT-6 Luna is $0.10/$0.50 — half this row on both sides — and is now OpenAI's recommended cost pick. Not deprecated, so nothing forces a move; measure before you migrate	down 80% from the $1.00/$0.10/$6.00 this board recorded on 2026-07-30; the long-context tier above 272K input tokens is $0.40/$0.04/$0.50/$1.80, now stated on the model's own page	src1 src2 src3 src4 2026-10-22
GPT-5.6 Sol OPENAI · PREVIOUS	$4.00	$0.4	**$20.00**	−50%	1.05M	128K	effort knob	Yes	Yes	Yes	both	Closed		The hardest reasoning and longest agent runs, where an eval already signed off on 5.6	Superseded twice over: GPT-6 Sol is $2.00/$10.00, half this row, and OpenAI's models page now recommends GPT-6 Sol where it used to name GPT-5.6. This row's $4.00/$20.00 is still promotional through at least 2026-11-21 — so the older model is both the dearer one and the one on a clock	cut on 2026-08-22 from $5.00/$0.50/$6.25/$30.00, and labelled promotional "at least through November 21, 2026". Above 272K input tokens the whole request bills at 2x input and 1.5x output — $8.00/$0.80/$10.00/$30.00 — now stated on the model's own page	src1 src2 src3 src4 2026-10-22
GPT-5.6 Terra OPENAI · PREVIOUS	$2.00	$0.2	**$12.00**	−50%	1.05M	128K	effort knob	Yes	Yes	Yes	both	Closed		The default working model — coding, tools, long context	OpenAI's models page no longer names Terra — its balance pick is now GPT-6 Sol, which is $2 cheaper on output at the same $2.00 input. There is no GPT-6 Terra. Not deprecated; re-run your eval before moving	long-context tier above 272K input tokens: $4.00/$0.40/$5.00/$18.00, now stated on the model's own page	src1 src2 src3 src4 2026-10-22

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-6 Astra OPENAI · LATEST	$10.00	$1.00	**$50.00**	−50%	1.05M	128K	effort knob	Yes	?	Yes	both	Closed		OpenAI's recommended starting point for the hardest reasoning and coding work	Ties Claude Fable 5.1 for the dearest output on the board at $50.00, and its $1.00 cache read is four times Fable 5.1's $0.25. Reasoning effort starts at low — there is no none on Astra, though GPT-6 Sol and Luna both have it — so every call pays for some thinking, at the output rate	above 272K input tokens the whole request bills at 2x input and cache and 1.5x output: $20.00/$2.00/$25.00/$75.00, and Batch halves every column. OpenAI's models page names this as the place to start for complex reasoning and coding	src1 src2 src3 src4 2026-10-22
GPT-6 Luna OPENAI · LATEST	$0.1	$0.01	**$0.5**	−50%	1.05M	128K	effort knob	Yes	?	Yes	both	Closed		OpenAI's pick for cost-sensitive, high-volume work — the cheapest OpenAI row on the board by half	Half GPT-5.6 Luna's $0.20/$1.20 on both sides, and GPT-5.6 Luna carries no deprecation date — so nothing forces the move, and nothing but your own eval tells you the cheaper row does the same job	above 272K input tokens: $0.20/$0.02/$0.25/$0.75 for the whole request. Batch and Flex halve every column · 922,000 max input tokens	src1 src2 src3 src4 src5 2026-10-22
GPT-6 Luna · Decisions API OPENAI · LATEST	$0.1	–	**free**	–	–	–	—	Yes	—	—	none	Closed		Typed answers — predicate, choice, score — over text and images from a frontier lab's own endpoint, with ZDR and HIPAA for eligible customers	Beta, with GPT-6 Luna the only model. At $0.10 per 1M input it is two and a half times Jev and five times Perplexity's Decider; OpenAI's "10x faster than the Responses API" is its own measurement. Images must be inline base64	the /v1/decisions rate for gpt-6-luna: input only, with no cache-read, cache-write or output charge; regional-processing premiums and long-context input multipliers still apply. Other gpt-6-luna requests bill at the GPT-6 Luna row's rates. Public beta from 2026-10-06 · the Decisions guide does not state a per-request input limit for /v1/decisions	src1 src2 2026-10-22

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-6 Sol OPENAI · PREVIOUS	$2.00	$0.2	**$10.00**	−50%	1.05M	128K	effort knob	Yes	?	Yes	both	Closed		Workloads that need reasoning effort none at GPT-6 pricing — the one thing GPT-6.1 Sol does not offer	Superseded on 2026-09-29: the model page now says "See GPT-6.1 Sol for the newer Sol model" and OpenAI's starting points list GPT-6.1 Sol instead. Not deprecated and no shutdown date, but its $0.20 cache read is twice GPT-6.1 Sol's $0.10 at the same input and output rates	above 272K input tokens the whole request bills at 2x input and cache and 1.5x output: $4.00/$0.40/$5.00/$15.00. Batch and Flex halve every column · 922,000 max input tokens	src1 src2 src3 src4 src5 2026-10-22
GPT-6.1 Sol OPENAI · LATEST	$2.00	$0.1	**$10.00**	−50%	1.05M	128K	effort knob	Yes	?	Yes	both	Closed		OpenAI's recommended balance of capability and cost from 2026-09-29, and the cheapest cache read of any OpenAI row relative to its input	Reasoning effort starts at low — none and minimal are not supported, where GPT-6 Sol offers none — so moving a no-thinking workload from GPT-6 Sol raises its output bill even at the same list price. Above 272K input every token in the request bills at the higher tier	released 2026-09-29. Above 272K input tokens the whole request bills at 2x input and cache and 1.5x output: $4.00/$0.20/$5.00/$15.00. Batch halves every column. Same input and output as GPT-6 Sol; the cache read is half, at 0.05x input · 1,050,000 window and 128,000 max output, both on the model's own page	src1 src2 src3 src4 src5 2026-10-22
Perplexity Decider Flash v1.2 PERPLEXITY · SONAR · LATEST	$0.01	—	**free**	—	—	—	—	Yes	—	—	unverified	Not verified		High-volume triage where v1.2's accuracy is unneeded and a cent per million input tokens is the point	The cheapest typed-answers row on the board; calibrate it on your own data before trusting the saving	input only, output free, no per-request fee — half the price of v1.2 · no input budget is stated on the v1.2 pages fetched this pass; v1.1 stated 262,144 tokens per request	src1 src2 2026-10-22
Perplexity Decider v1.2 PERPLEXITY · SONAR · LATEST	$0.02	—	**free**	—	—	—	—	Yes	—	—	unverified	Not verified		Typed decisions over text, JSON or images at two cents per million input tokens	Perplexity does not say whether the weights are published. Aliases bill at the rate of the model they select	input only, output free, no per-request fee. Supersedes v1.1: the quickstart prices v1.2 and flash-v1.2 only · no input budget is stated on the v1.2 pages fetched this pass; v1.1 stated 262,144 tokens per request	src1 src2 2026-10-22

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 PERPLEXITY · SONAR · LATEST	$0.25	$0.062	**$2.50**	—	128K	—	—	—	—	—	unverified	Closed		Search-grounded answers with citations, no retrieval stack to build	Search is now billed per tool call, not per request: $0.0025 per web search is $2.50 per 1,000, below Chat Completions' $5–$12 request fee — but an agent that searches three times pays three. Output rose from $1.00 to $2.50, so long answers cost more than they did	Agent API rate for perplexity/sonar, the endpoint Perplexity supports since Sonar Chat Completions support ended on 2026-09-27; plus $0.0025 per web-search call and $0.0005 per URL fetch. Flex bills 0.5x and priority 2x. The pricing page still lists Chat Completions Sonar at $1/$1 plus $5–$12 per 1,000 requests, and Chat Completions requests are being reformulated as Agent API requests — Perplexity does not say which rate those bill at · 128K was stated on the Sonar model page, which became the Agent API migration guide on 2026-09-27; no page restates it	src1 src2 src3 2026-10-22
Sonar Pro PERPLEXITY · SONAR · DEPRECATED	$3.00	—	**$15.00**	—	200K	—	—	—	—	—	none	Closed		Deeper grounded research with more sources per answer	Sonar Chat Completions support ended on 2026-09-27. The Agent API has no Sonar Pro model — Perplexity's migration guide maps it to the `fast` preset — and asynchronous requests are no longer supported. Synchronous and streaming calls keep working while they are reformulated as Agent API requests, rolling out by model, and Perplexity does not say at which rate. The price shown is the Chat Completions rate the pricing page still lists	plus $6–$14 per 1,000 requests · stated on the model's own page; not restated since the model page became the migration guide on 2026-09-27	src1 src2 2026-10-15

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 Reasoning Pro PERPLEXITY · SONAR · DEPRECATED	$2.00	—	**$8.00**	—	128K	—	always on	—	—	—	none	Closed		A reasoning-capable grounded tier between Sonar and Sonar Pro	Sonar Chat Completions support ended on 2026-09-27. The Agent API has no Sonar Reasoning Pro model — Perplexity's migration guide maps it to the `low` preset — and asynchronous requests are no longer supported. Synchronous and streaming calls keep working while they are reformulated as Agent API requests, rolling out by model, and Perplexity does not say at which rate. The price shown is the Chat Completions rate the pricing page still lists	plus $6/$10/$14 per 1,000 requests by low/medium/high search-context size — the same request-fee schedule as Sonar Pro · stated on the model's own page; not restated since the model page became the migration guide on 2026-09-27	src1 src2 2026-10-15
Jev 1.13 TYPESAFE AI · PREVIOUS	$0.042	—	**free**	—	64K	—	—	—	—	—	unverified	Closed		Routing, triage, guardrails and rubric scoring where code needs a typed answer and a calibrated probability, not prose	Early access with limits that "can change without notice", English-first, text only, and no fine-tuning. TypeSafe's speed and cost multiples are measured on its own workflows. Also served through Vercel's and Netlify's AI gateways	billed per input token only — output is free, because a System One model generates no text. $42 per billion input tokens. Not comparable per token with the language-model rows above: it answers typed questions, it does not write · 64K per request for state plus all questions; 32K for state plus the single longest question. Vercel's AI Gateway lists 32,000	src1 src2 src3 2026-10-22

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
Jev 1.14 TYPESAFE AI · LATEST	$0.042	—	**free**	—	64K	—	—	—	—	—	unverified	Closed		Routing, triage, guardrails and rubric scoring where code needs a typed answer and a calibrated probability, not prose	Pin jev-1.14.0 rather than jev-latest if you have tuned confidence thresholds. Early access with limits that “can change without notice”, English-first, text only, and no fine-tuning	billed per input token only — output is free, because a System One model generates no text. $42 per billion input tokens. Not comparable per token with the language-model rows above: it answers typed questions, it does not write · 64K per request for state plus all questions; 32K for state plus the single longest question. An opt-in long-context header raises both to 262K at the same per-token price	src1 src2 src3 2026-10-22
Grok 4.3 XAI · GROK · PREVIOUS	$1.25	$0.2	**$2.50**	−50%	1M	—	effort knob	Yes	?	Yes	automatic	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 src2 src3 2026-10-22
Grok 4.5 XAI · GROK · PREVIOUS	$2.00	$0.3	**$6.00**	—	500K	—	effort knob	Yes	?	Yes	automatic	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 src2 src3 2026-10-22
Grok 4.6 XAI · GROK · PREVIOUS	$2.00	$0.5	**$6.00**	—	500K	—	effort knob	Yes	—	Yes	automatic	Closed		Long-running agents and codebase work under a 200K prompt	Superseded by Grok 4.7 at the identical rate, so there is no price reason to stay. The cache discount is still weaker than 4.5's — a prefix-heavy loop may 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 · xAI states "no text output limit" rather than a number	src1 src2 src3 2026-10-22

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 4.7 XAI · GROK · LATEST	$2.00	$0.5	**$6.00**	—	500K	—	effort knob	Yes	—	Yes	automatic	Closed		xAI's frontier row for coding, agents and knowledge work, at Grok 4.6's price	Batch API is not supported on this row, so the 20% batch discount Grok 4.3 carries does not apply. A Grok 4.7 Fast variant at twice the rate exists, but only through Cursor and Grok Build — not on the public API	at/above 200K input: $4.00 in / $1.00 cached / $12.00 out, applied to every token in the request. Identical to Grok 4.6. A US regional endpoint bills 1.1x · xAI states "no text output limit" rather than a number	src1 src2 src3 src4 2026-10-22
Grok Build 0.1 XAI · GROK · LATEST	$1.00	$0.2	**$2.00**	—	256K	—	not stated	Yes	?	Yes	automatic	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 src2 src3 2026-10-22
GLM-4.7 Z.AI · GLM · PREVIOUS	$0.6	$0.11	**$2.20**	—	200K	128K	effort knob	—	—	Yes	automatic	Open	MIT (reported)	Well-priced open-weight middle ground	Superseded by 4.7-FlashX at a tenth the input price for lighter work	stated on the model guide	src1 src2 src3 2026-10-22
GLM-4.7-FlashX Z.AI · GLM · LATEST	$0.07	$0.01	**$0.4**	—	200K	128K	effort knob	—	—	Yes	automatic	Open	MIT (reported) — no Hugging Face repo for FlashX itself; GLM-4.7-Flash is MIT	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 · stated on the model guide	src1 src2 src3 2026-10-22
GLM-5 Z.AI · GLM · PREVIOUS	$1.00	$0.2	**$3.20**	—	200K	128K	effort knob	—	—	Yes	automatic	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	stated on the model guide	src1 src2 src3 2026-10-22
GLM-5.2 Z.AI · GLM · PREVIOUS	$1.40	$0.26	**$4.40**	—	1M	128K	effort knob	—	—	Yes	automatic	Open	MIT — stated on the Hugging Face model card	MIT-licensed frontier-adjacent coding model you can self-host	Superseded by GLM-5.3 on 2026-09-03 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 MIT, GLM-5.3's carry a revenue-gated security-review clause	both figures stated on the GLM-5.2 guide	src1 src2 src3 src4 2026-10-22

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-5.3 Z.AI · GLM · LATEST	$1.40	$0.26	**$4.40**	—	1M	128K	always on	—	—	**Yes**	automatic	Open	GLM-5.3 License — MIT-style grant, plus a Z.ai security review before commercial use by a MaaS business above $10B revenue over 12 months	The current GLM flagship, with a first-party 1M window on the record	Read the licence before you ship. The GLM-5.3 License is an MIT-style grant with one extra condition: a model-as-a-service business with more than $10B revenue over 12 months must pass a Z.ai security review before commercial use. GLM-5.2 and GLM-5.3-Flash are plain MIT. 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	src1 src2 src3 src4 2026-10-22
GLM-5.3-Flash Z.AI · GLM · LATEST	$0.15	$0.03	**$0.5**	—	1M	128K	always on	**Yes**	?	**Yes**	automatic	Open	MIT — stated in the Hugging Face LICENSE	The cheapest tier of the GLM-5.3 generation	The launch promotion has ended, so $0.15/$0.50 is now the rate you are billed. Unlike GLM-5.3, its weights are plain MIT on Hugging Face — and it takes images and video, which GLM-5.3 does not. Thinking cannot be switched off	list rate. The launch promotion that halved it to $0.075/$0.015/$0.25 ended on 2026-09-09 as scheduled, and is no longer on the pricing page · stated on the GLM-5.3-Flash guide	src1 src2 src3 src4 2026-10-22

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-V4.1-Flash ON BY DEFAULT AND BEST-EFFORT; A CACHE HIT IS $0.006 AGAINST A $0.30 MISS AT PEAK (2%), THE CHEAPEST CACHE READ ON THIS BOARD. FIM COMPLETION IS IN BETA AND WORKS IN NON-THINKING MODE ONLY	automatic	—	no premium	$0.162	**$0.030**	82%
GPT-5.6 Terra CACHED INPUT AT 10% OF BASE; IMPLICIT MODE PLACES THE BREAKPOINT FOR YOU, EXPLICIT-ONLY MODE LETS YOU PLACE IT, AND AN ENTRY LIVES AT LEAST 30 MINUTES AFTER ITS LAST WRITE OR REUSE	both	30m	$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	automatic	5m / 1h	$3.00	$1.650	**$0.435**	74%
Mistral Medium 3.5 CACHED INPUT BILLS AT 10% OF INPUT IN 64-TOKEN BLOCKS, SO A PROMPT UNDER 64 TOKENS NEVER HITS; A PROMPT_CACHE_KEY RAISES THE CHANCE OF A HIT BUT DOES NOT GUARANTEE ONE. WHETHER CACHING APPLIES WITHOUT THE KEY, AND THE TTL, ARE NOT STATED	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 this panel applies a cache only where the provider says how it works. Mistral's docs now give the rate, $0.15 cached, but say a `prompt_cache_key` raises the chance of a hit without guaranteeing one, and state no TTL, so the Mistral row is priced here as if nothing hits. Read that 0% as unassumed, not as absent. The inversion this panel exists to show is real — this row just cannot demonstrate it until Mistral says how its cache behaves.

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, and a 50% Batch API discount on every listed model. source https://developers.openai.com/api/docs/pricing

BENEFIT Prefix caching at 10% of the input rate, in two modes: implicit, which places the breakpoint for you, or explicit-only, where you mark it. On GPT-5.6 and later a prefix must reach 1,024 tokens before it caches, and an entry lives at least 30 minutes after its last use. source https://developers.openai.com/api/docs/guides/prompt-caching

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 on GPT-5.6, GPT-6 Sol and GPT-6 Luna — GPT-6 Astra and GPT-6.1 Sol start at low, with no none — 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/gpt-6.1-sol

BENEFIT The Decisions API (POST /v1/decisions, public beta from 2026-10-06) returns typed answers — a probability, a choice from your set, or a rubric score — from gpt-6-luna at $0.10 per 1M input tokens with no output charge. It supports ZDR and HIPAA for eligible customers, and US and EEA regional processing. source https://developers.openai.com/api/docs/guides/decisions

---

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 — 0.05x on Claude Opus 5.5 and, since 2026-10-07, Claude Sonnet 5.5, and 0.025x on Claude Fable 5.1 and Mythos 5.1. 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

WATCH Claude Haiku 5.5 is the first Haiku billed by prompt length: $0.10/$0.50 up to 100,000 input tokens, $0.50/$2.50 above, applied to the whole request. Anthropic says around 90% of requests to the previous Haiku fall below the line; check where yours fall before you budget the low rate. source https://platform.claude.com/docs/en/models/haiku-5-5/overview

BLOCKER Claude Fable 5 and now Claude Opus 5.5 ship 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 will treat a refusal as an answer. Opus 5.5 adds a biology classifier and a `reasoning_extraction` category. source https://platform.claude.com/docs/en/models/opus-5-5/overview

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 5.5 shipped on 2026-09-28 at Sonnet 5's $2/$10, and that is the standard rate — the rise to $3/$15 once announced for 2026-09-01 was withdrawn. 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

WATCH Gemini 4 Argon (GA 2026-10-13) bills $2/$10 introductory through 2026-11-30, then $4/$20 from 2026-12-01, with cached input doubling $0.10 to $0.20 alongside. It is paid-tier only with implicit caching exclusively — no free tier, no explicit caches — on a 1,048,576-token unified input-plus-output budget. 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 ($0.50–$4.50 per 1M tokens per hour — $0.50 on Gemini 3.8 Flash until it doubles on 2027-01-01) 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

WATCH Implicit caching is on by default for Gemini 2.5 and newer, but only above a minimum prompt: 4,096 tokens on the 3.x Flash rows and 3.1 Pro Preview, 2,048 on 2.5 Flash. A system prompt under that line caches nothing, and nothing tells you. source https://ai.google.dev/gemini-api/docs/caching

WATCH The Gemini API now runs a Contributor Tier, invite-only private preview, where content may be used to improve Google's products — the same data-for-discount shape as Meta's contributor mode, but not yet a public rate card. 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 rows. 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 5.3 and 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 models 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://docs.mistral.ai/inference/pricing

BENEFIT Mistral Large 4 (public preview 2026-10-06): 52B active of 1.05T parameters, 1M window, docs badge Open, at $1.36 in / $0.14 cached / $4.18 out on Mistral's own pricing page. The weights ship end-October with an unread licence — until then it is an API-only row. source https://docs.mistral.ai/inference/pricing

---

Mitigation: Price a prefix-heavy Mistral workload on the per-model cached rate, then measure your real hit rate — Mistral's own docs say the cache key raises the odds of a hit rather than guaranteeing one.

DeepSeek

BLOCKER Names now resolve to checkpoints, not models: since 2026-10-17 deepseek-flash is served by DeepSeek-V4.1-Flash-1017, and the legacy deepseek-v4-flash names are retired but still answer from 1017 at the Flash price. deepseek-v4-pro still does not route elsewhere. source https://api-docs.deepseek.com/quick_start/pricing

BLOCKER Peak/off-peak billing carries over to V4.1-Flash: $0.30 in / $1.20 out at peak, $0.15 / $0.60 off-peak. Peak hours are 01:00–04:00 and 06:00–10:00 UTC, Monday through Friday, excluding Chinese public holidays — which count as off-peak in full, as do weekends. So a nightly batch job scheduled inside one pays double. source https://api-docs.deepseek.com/quick_start/pricing

WATCH GPT-6 Luna, new this week at $0.10/$0.50, undercuts V4.1-Flash on input and output at every hour — V4.1-Flash is $0.15/$0.60 off-peak and $0.30/$1.20 at peak. DeepSeek still wins on cache hits, $0.003–$0.006 against $0.01, so a cache-heavy workload can still come out cheaper here. Price it on your own hit rate. source https://api-docs.deepseek.com/quick_start/pricing

BENEFIT The most aggressive cache discount on the board: a V4.1-Flash cache hit is 2% of a miss ($0.006 against $0.30 at peak), against the 10% standard elsewhere. Caching is on by default and best-effort. source https://api-docs.deepseek.com/guides/kv_cache/

BENEFIT V4.1-Flash's weights are on Hugging Face under the MIT License, stated on the model card itself — self-hosting and fine-tuning carry no commercial restriction. source https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash

WATCH The pricing page states rates "may vary and DeepSeek reserves the right to adjust them" — and V4.1-Flash arrived at new rates the same day it launched. 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

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: Pin deepseek-flash rather than a legacy flash name; treat deepseek-v4-pro as still live at its own peak/off-peak rates until DeepSeek says otherwise; price 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/developers/pricing

WATCH Grok 4.7 (2026-09-21) and Grok 4.6 bill the same $2/$6 with cached input at $0.50 — up from Grok 4.5's $0.30. On a prefix-heavy agent loop the newer models are the more expensive ones; check your cache-hit rate before migrating on the version number. source https://docs.x.ai/developers/pricing

WATCH Every model bills its long-context rate on every token once the prompt reaches 200K: grok-4.7 and grok-4.6 go from $2/$6 to $4/$12. The advertised 500K–1M window costs 2x per token to actually fill. source https://docs.x.ai/developers/pricing

WATCH Grok 4.7 and 4.6 state "no text output limit" rather than a max-output number, and Batch is not supported on either — only grok-4.3 carries the 20% batch discount. Caching is automatic on every Grok language model. source https://docs.x.ai/developers/advanced-api-usage/prompt-caching

---

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

BENEFIT Caching is automatic on all three models, and cheapest on K3 — hits at 10% of input ($0.30 against $3.00), against about 17% on K2.6 and 20% on K2.7 Code ($0.16 and $0.19 against $0.95). There is also a dedicated cheaper coding model, K2.7 Code at $0.95/$4.00 with a 262,144-token context. source https://platform.kimi.ai/docs/pricing/chat

WATCH K3 ships under its own Kimi K3 License: a model-as-a-service operator above $20M revenue over any 12 months needs a separate agreement, and on-screen credit applies above 100M MAU or $20M monthly revenue. K2.7 Code and K2.6 stay on the Modified MIT licence. Fine for most teams; read it if you are either. source https://huggingface.co/moonshotai/Kimi-K3

BENEFIT All three models are priced on one first-party page, and the Batch API bills K2.7 Code and K2.6 at 60% of the standard rate. K3 is not offered on batch, but its cache write is now priced: $3.00 at the default 5-minute TTL, $6.00 for an hour. source https://platform.kimi.ai/docs/pricing/batch

WATCH The built-in $web_search tool is expected to be deprecated on 2026-10-20; Kimi's docs already advise against it. A grounded-answer pipeline built on it needs its own retrieval, or the new /v1/tools/search API at $0.002 per call. source https://platform.kimi.ai/docs/pricing/websearch

---

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 GLM-5, GLM-5.2 and GLM-5.3-Flash are MIT on Hugging Face — unrestricted commercial use if you self-host. GLM-5.3 is the exception: an MIT-style grant with a security-review clause for model-as-a-service businesses above $10B revenue. source https://huggingface.co/zai-org/GLM-5.3-Flash

WATCH The pricing page states no context windows — the model guides do: 1M in / 128K out on GLM-5.3, 5.2 and 5.3-Flash, 200K / 128K on GLM-5, 4.7 and 4.7-FlashX. Caching is implicit on all mainstream models, with no flag to set. source https://docs.z.ai/guides/capabilities/cache

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: Size prompts against the model guide, not the pricing page — only the guides state a window.

Alibaba · Qwen

BENEFIT Model Studio documents caching as a multiplier on most models — explicit cache hits at 10% of the input rate, creation at 125% — but not on the qwen3.8 line, whose cached rates are printed on each model's own page instead. 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 open Qwen lines are Apache-2.0 on their Hugging Face cards — Qwen3.5-397B-A17B among them — the least legally ambiguous open weights on this board for a commercial product. source https://huggingface.co/Qwen/Qwen3.5-397B-A17B

WATCH Max is a closed API product. Qwen3.8's open checkpoint exists, but under a custom licence of its own name and text-only — the Max API's vision and video are not in the weights. Do not plan a self-hosting story around the Max API's feature set. source https://huggingface.co/Qwen/Qwen3.8-2.4T-A95B

WATCH First-party and host prices diverge — Qwen3.7-Max is $2.50 per 1M input on Model Studio against $1.50 on Together, and Together moved that rate up from $1.25 this week. Shop it before you commit, and re-check it. 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 — for its closed models: Muse Spark and, new in preview, Muse Wonder 1.0 at $4.00 in / $0.20 cached / $20.00 out with no contributor tier. source https://dev.meta.ai/docs/pricing-rate-limits.md

BENEFIT Muse Spark bills one flat rate 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://dev.meta.ai/docs/pricing-rate-limits.md

BENEFIT Muse Wonder 1.0 (preview) is Meta's most capable model for complex software engineering and long-horizon agents, with text, image, video, audio and PDF input, a 1,048,576-token window and reasoning effort from minimal to max. Muse Wonder Full, the higher-capability variant, is Muse Code subscription only. source https://dev.meta.ai/docs/models.md

BENEFIT Muse Spark Open is the open-weight release of Muse Spark — 816B parameters, 42B active — with no Meta endpoint; Together serves it at $1.25/$0.15/$4.25, the same rate as closed Spark on Meta's API, so the choice is custody and licence, not price. source https://www.together.ai/models/muse-spark-open

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's 10M-token context window is stated by Meta's own Llama 4 model card — by a wide margin the largest on this board, open-weight or otherwise. What it lacks is hosts: Together lists it only for fine-tuning and Groq deprecated it for free and developer tiers. source https://raw.githubusercontent.com/meta-llama/llama-models/main/models/llama4/MODEL_CARD.md

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

WATCH Contributor limits now read 150 RPM against the standard tier's 3,000 (this board said 100 until this pass), and Muse Code subscriptions default to contributor mode — the data-sharing tier is the default there, not the opt-in. source https://dev.meta.ai/docs/pricing-rate-limits.md

---

Mitigation: Treat Muse Spark, Muse Wonder and Llama as three unrelated products. For Llama and Spark Open, pick the host first, then price the workload; for Spark and Wonder, read Meta's own terms — the Llama licence applies to neither.

MiniMax

BENEFIT M3.1 Flash bills $0.15 in / $0.60 out with $0.015 cache reads — half of M3's $0.30/$1.20, and a quarter of its $0.06 cache read. Weights are on Hugging Face under the minimax-community licence; the licence file was not re-read this pass. source https://platform.minimax.io/docs/guides/pricing-paygo

BENEFIT M3 lists at $0.30 in / $1.20 out with cached reads at $0.06 — an 80% discount — on MiniMax's own pricing page, and caching is automatic from 512 input tokens. source https://platform.minimax.io/docs/guides/pricing-paygo

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 labelled "Permanent 50% off" from $0.60/$2.40 below 512K. Budget the struck-through rate if a word on a pricing page is not a contract to you. source https://platform.minimax.io/docs/guides/pricing-paygo

WATCH The weights are open under the MiniMax Community License, which grants non-commercial use by default: commercial use requires "Built with MiniMax M3" attribution, a notice below $20M yearly revenue and written authorisation above it, and it prohibits military use. source https://huggingface.co/MiniMaxAI/MiniMax-M3

Mitigation: Read the Community License before you self-host commercially; the hosted API is priced first-party.

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/docs/getting-started/pricing

BLOCKER Search is priced apart from tokens, and it usually dominates. On the Agent API it is per tool call — $0.0025 per web search, $0.0005 per URL fetch — so an agent that searches five times pays five. On the Chat Completions endpoint it was $5–$14 per 1,000 requests by search-context size. Model the searches first; the token rate is the small number. source https://docs.perplexity.ai/docs/agent-api/models

WATCH Sonar Chat Completions support ended on 2026-09-27. Synchronous and streaming requests keep working while they are reformulated as Agent API requests, rolling out by model; asynchronous requests are no longer supported. Sonar Pro and Sonar Reasoning Pro have no Agent API model of their own — they map to the fast and low presets, which run other labs' models at those models' rates. source https://docs.perplexity.ai/docs/sonar/models/sonar

BENEFIT The Decisions API (POST /v1/decisions, pplx-decider-v1.2 and flash-v1.2) answers yes/no, choice and scored questions over text, JSON or images with probabilities instead of text, at $0.02 and $0.01 per 1M input tokens with output free. source https://docs.perplexity.ai/docs/decisions/quickstart

---

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 560 on Llama 3.1 8B. If time-to-first-token is the product, this is the lever. source https://console.groq.com/docs/models

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

WATCH Most of the menu caps context at 131,072 tokens — MiniMax M2.7, at 196,608, is the exception — well below the 1M windows available elsewhere. source https://console.groq.com/docs/models

WATCH Both Llama rows read "Contact Sales" because Groq deprecated them for free and developer tiers on 2026-08-16; only enterprise customers on a committed-spend contract still get them. groq.com/pricing now redirects to the homepage. The rows below keep their speeds and drop the prices this board had carried since 2026-08-06. source https://console.groq.com/docs/deprecations

---

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 carry dated labels, and the date is the model. Together still serves 'DeepSeek V4 Flash 0731' at $0.14/$0.03/$0.28 — a checkpoint DeepSeek retired on its own API on 2026-09-10. Pin the exact model string, not the family name, or you will not know which model you are billed for. 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.

TypeSafe AI

BENEFIT Jev 1.14 (2026-10-14) supersedes 1.13 at the same $0.042 per 1M input tokens, with an opt-in long-context header raising both context limits to 262K at the same per-token price. source https://docs.typesafe.ai/models

BENEFIT Jev is not a language model: a request is a block of state plus typed questions — Choice, Score or a yes/no "Noul" — and every answer comes back as a probability distribution with a confidence, never as text. All questions on one state are evaluated in a single parallel pass. source https://docs.typesafe.ai/introduction

BENEFIT Billed on input only: $0.042 per 1M tokens ($42 per billion), and output is free because nothing is generated. Packing many questions onto one state is the documented way to cut cost further; there is no batch or caching discount. source https://docs.typesafe.ai/models

WATCH Early access, and the limits say so: 250,000 tokens/s and 1,200 requests/min, which TypeSafe says "can change without notice" while demand is high. English is where accuracy is best; other languages, CJK included, are handled "not equally well". source https://docs.typesafe.ai/models

WATCH TypeSafe says of its own price, "We can't prove it isn't subsidized." Its speed and cost multiples against frontier LLMs come from its own workflows — a claim to test on yours, not a result this board has measured. source https://typesafe.ai/blog/introducing-system-one-models-and-jev

BENEFIT Not trained on customer requests or responses, and the same weights serve every account — there is no fine-tuning; domain rules go in the state and the question instructions. Zero data retention is offered to enterprise customers. source https://docs.typesafe.ai/models

---

Mitigation: Pin jev-1.14.0 rather than jev-latest if you have tuned confidence thresholds, and run your own decisions through it and through the LLM call it would replace before switching.

milliseconds.ai

BENEFIT decision-machine-1 bills $0.04 per 1M input tokens with output free, and test keys carry 125M free input tokens a month — enough to evaluate a real workload before paying. source https://www.milliseconds.ai/pricing/

WATCH A request's text is capped at 20,000 characters (or up to 32 texts in a batch), and production rate limits are spend-gated: 1,000 requests/min at Tier 1, rising with lifetime paid spend and time since first payment. source https://www.milliseconds.ai/pricing/

WATCH It serves a System One-compatible route alongside its own endpoints, so TypeSafe's SDK can point at it — but its terms restrict using the service to benchmark a competing model. Read them before you run a bake-off. source https://www.milliseconds.ai/legal/terms/

Mitigation: Evaluate on test keys first; check the terms before publishing any comparison.

meraGPT

BENEFIT Decider 1 is $0.04 per 1M input tokens — up from $0.03 last pass — with output at $0.00, speaks the System One schema — the TypeSafe SDK works by changing its base URL — and states that the same input returns the same output. source https://meragpt.com/models/state-decider-1

WATCH Much tighter than Jev: 4,096 tokens per request for state and questions together, and at most ten options per choice question. Text only. source https://meragpt.com/models/state-decider-1

WATCH The comparison numbers on its page are its own single run, scored against a reference built from an ensemble of teacher models — the page says so itself. New on 2026-09-22; weights proprietary. source https://meragpt.com/models/state-decider-1

Mitigation: Split large label sets into groups with a "none of these" option, as meraGPT's own page advises, and keep state under the 4,096-token cap.

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 2026-10-22

Speed is the product — up to 1,000 tokens/sec on a narrow menu, mostly at a 131,072-token context cap. console.groq.com/docs/models carries prices, contexts and speeds; both Llama rows are now enterprise-only and priced "Contact Sales", so they show speed without a rate.

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	—	—	—	131K	560 tok/s	enterprise-only since 2026-08-16 (deprecated for free and developer tiers), priced "Contact Sales". The $0.05/$0.08 this board carried until 2026-09-17 was dropped this pass: it is a rate for a tier that no longer exists publicly. 560 tok/s re-read this pass
Llama 3.3 70B Versatile	—	—	—	131K	280 tok/s	enterprise-only since 2026-08-16, priced "Contact Sales" — the carried $0.59/$0.79 was dropped this pass. Together lists this model at $1.04/$1.04. 280 tok/s re-read this pass
Qwen3.8-27B	$0.8	—	$4.00	131K	450 tok/s	preview model; replaced Qwen 3.6 27B, which Groq deprecated on 2026-09-14. Context corrected this pass from a mistyped 131,042 to the 131,072 the page states; max completion 16,384. Alibaba's own rate for this model is $0.50/$3.00

TOGETHER AI

VERIFIED 2026-10-22

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 peak-hour rate on in/out exactly, at 3x DeepSeek's peak cached-hit rate ($0.044). The legacy 'DeepSeek V4 Pro' row at $1.74/$0.20/$3.48 is gone from Together's table this pass
DeepSeek V4.1 Flash	$0.3	$0.006	$1.20	—	—	matches DeepSeek's own peak-hour rate on in/out and on cache — the first host row this board has seen that does
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 — the V4-Flash checkpoint DeepSeek itself retired on 2026-09-10. Half DeepSeek's current V4.1-Flash off-peak input: cheap because it is an older model, which is the trap this table exists to show
Kimi K3	$3.00	$0.3	$15.00	—	—	matches first-party (platform.kimi.ai)
Qwen3.7-Max	$1.50	$0.3	$4.50	—	—	up this pass from $1.25/$0.13/$3.75 — now 60% of the first-party Model Studio rate, not half
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	—	—	Groq still lists this model, but enterprise-only and priced "Contact Sales" — so this is the only public per-token rate on the board for it
Muse Glimmer 30B	$0.35	$0.04	$1.50	—	—	Meta's first Apache-2.0 model; no first-party endpoint to compare against
Qwen3.8 Flash	$0.09	—	$0.28	—	—	new this pass — below Alibaba's own $0.15/$0.47 for the same model

HUGGING FACE **VERIFIED 2026-10-15**

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.

Claude Sonnet 5.5
anthropic

10–01 **$10.00** → 10–08 **$10.00** → 10–15 **$10.00** → 10–22 **$10.00**

VENDOR CHANGE 2026-10-08 · cached_in — $0.20 → $0.10 (0.1x → 0.05x input): Anthropic's release note of 2026-10-07 halves the Sonnet 5.5 cache read, and the pricing page's multiplier section states $0.10. The same page's model table still prints $0.20; the row follows the dated change.

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** → 09–10 **$1.20** → 09–17 **$1.20** → 09–24 **$1.20** → 10–01 **$1.20** → 10–08 **$1.20** → 10–15 **$1.20** → 10–22 **$1.20**

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.

VENDOR CHANGE 2026-09-10 · in, cached_in, out — DeepSeek retired V4-Flash on 2026-09-10. The deepseek-v4-flash name is still accepted but is served by DeepSeek-V4.1-Flash and billed at its price, so the row moves from V4-Flash's own

$0.44/$0.014/$1.32 at peak to the $0.30/$0.006/$1.20 the name now bills. The row is marked deprecated.

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** → 09–10 **$3.96** → 09–17 **$3.96** → 09–24 **$3.96** → 10–01 **$3.96** → 10–08 **$3.96** → 10–15 **$3.96** → 10–22 **$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** → 09–10 **$3.75** → 09–17 **$3.75** → 09–24 **$3.75** → 10–01 **$3.75** → 10–08 **$3.75** → 10–15 **$3.75** → 10–22 **$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** → 09–10 **$1.20** → 09–17 **$1.20** → 09–24 **$1.20** → 10–01 **$1.20** → 10–08 **$1.20** → 10–15 **$1.20** → 10–22 **$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** → 09–10 **$20.00** → 09–17 **$20.00** → 09–24 **$20.00** → 10–01 **$20.00** → 10–08 **$20.00** → 10–15 **$20.00** → 10–22 **$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

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

openai

08–27 **$12.00** → 09–03 **$12.00** → 09–10 **$12.00** → 09–17 **$12.00** → 09–24 **$12.00** → 10–01 **$12.00** → 10–08 **$12.00** → 10–15 **$12.00** → 10–22 **$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** → 09–10 **$4.00** → 09–17 **$4.00** → 09–24 **$4.00** → 10–01 **$4.00** → 10–08 **$4.00** → 10–15 **$4.00** → 10–22 **$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** → 09–10 **$3.60** → 09–17 **$3.60** → 09–24 **$3.60** → 10–01 **$3.60** → 10–08 **$3.60** → 10–15 **$3.60** → 10–22 **$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** → 09–10 **$7.50** → 09–17 **$7.50** → 09–24 **$7.50** → 10–01 **$7.50** → 10–08 **$7.50** → 10–15 **$7.50** → 10–22 **$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** → 09–10 **$1.60** → 09–17 **$1.60** → 09–24 **$1.60** → 10–01 **$1.60** → 10–08 **$1.60** → 10–15 **$1.60** → 10–22 **$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.

Sonar

07–30 **$1.00** → 08–06 **$1.00** → 08–14 **$1.00** → 08–20 **$1.00** →

perplexity

08–27 **$1.00** → 09–03 **$1.00** → 09–10 **$1.00** → 09–17 **$1.00** → 09–24 **$1.00** → 10–01 **$2.50** → 10–08 **$2.50** → 10–15 **$2.50** → 10–22 **$2.50**

VENDOR CHANGE 2026-10-01 · in, cached_in, out — $1.00 / — / $1.00 → $0.25 / $0.0625 / $2.50: Perplexity ended Sonar Chat Completions support on 2026-09-27 and moved Sonar to the Agent API, which prices perplexity/sonar at its own rate and bills search per tool call. The Chat Completions rate is still printed on the pricing page for requests being reformulated.

Decider 1
meragpt

09–24 **$0.00** → 10–01 **$0.00** → 10–08 **$0.00** → 10–15 **$0.00** → 10–22 **$0.00**

VENDOR CHANGE 2026-10-08 · in — $0.03 → $0.04 per 1M input tokens, as stated on meraGPT's Decider 1 page this pass. Output stays $0.00.

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: The contributor discount is real — if your repo already is.

2026-10-22

Gemini 4 Argon generally available at an introductory $2/$10 that doubles on 1 December

Google's first Gemini 4 model left cyber-defender early access on 2026-10-13 for the paid API, AI Studio and Pro/Ultra subscribers. The intro rate ($2 in / $10 out / $0.10 cached) runs to 2026-11-30, then $4/$20/$0.20 — a dated doubling, not a cut to anchor on. Paid tier only, implicit caching exclusively, on a 1M unified input-plus-output budget with 65K default max output (1M with long decode). Google positions it for long-horizon agents; every benchmark on the launch post is Google's own. source https://ai.google.dev/gemini-api/docs/pricing

---

2026-10-22

Meta previews Muse Wonder 1.0 and its open-weight Spark escapes onto hosts

Meta's pricing page now lists a second closed model: Muse Wonder 1.0 in preview at $4/$0.20/$20 with no contributor tier, aimed at complex software engineering and long-horizon agents. In parallel, Muse Spark Open — the 816B-parameter open release of Spark — is served by Together at $1.25/$0.15/$4.25, exactly closed Spark's rate. And the contributor small print moved: contributor RPM now reads 150 (this board said 100), and Muse Code subscriptions default to contributor mode. source https://dev.meta.ai/docs/models.md

---

2026-10-22

MiniMax M3.1 Flash goes GA at $0.15/$0.60 — half of M3

MiniMax's own pricing page lists M3.1 Flash at $0.15 in / $0.015 cached / $0.60 out — half of M3 in and out, a quarter on cache reads — with no priority tier. Secondary coverage dates GA to 2026-10-19; the pricing page carries no date. The weights followed on Hugging Face under the minimax-community licence; read the licence file before self-hosting, and treat the capability columns as unstated rather than absent. source https://platform.minimax.io/docs/guides/pricing-paygo

---

2026-10-22

Mistral Large 4 previews at $1.36/$4.18 with open weights pending

Mistral's trillion-parameter flagship (52B active, 1M window, multimodal with function calling and agents) has been callable since 2026-10-06 at $1.36 in / $0.14 cached / $4.18 out on its own pricing page. The docs badge says Open but the weights land end-October with an unread licence — price the API, not self-hosting, and re-read the licence file before building on the weights. source https://docs.mistral.ai/inference/pricing

---

2026-10-22

Perplexity Decisions moves to v1.2 with a $0.01 Flash tier; DeepSeek swaps in the 1017 checkpoint

Perplexity's quickstart now prices pplx-decider-v1.2 at $0.02 and flash-v1.2 at $0.01 per 1M input, output free — v1.1 has no stated rate left, so the board follows the quickstart. DeepSeek's deepseek-flash is served by the V4.1-Flash-1017 checkpoint since 2026-10-17 at unchanged rates, with legacy flash names retired but still answering at the Flash price. source https://docs.perplexity.ai/docs/decisions/quickstart

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

- **Rows carried at 2026-10-15** — carried at 2026-10-15 verification, not re-read 2026-10-22: Kimi K2.7 Code — pricing tables render client-side; Together lists only dedicated tiers; Kimi K2.6 — pricing tables render client-side; Together lists only dedicated tiers; Llama 4 Scout — no first-party endpoint; Together's Scout listing is ambiguous across service tiers; Gemini 3.7 Flash — no longer on the Gemini pricing page; delisted without notice; Gemini 3.5 Flash — no longer on the Gemini pricing page; delisted without notice; Devstral 2 — retired; absent from the Mistral pricing page; Sonar Pro — deprecated row; its model page is now the Agent API migration guide with no rate table; Sonar Reasoning Pro — deprecated row; its model page is now the Agent API migration guide with no rate table
- **Groq — Llama per-model rates** — closed as a pricing question, not as a gap: console.groq.com/docs/deprecations says llama-3.1-8b-instant and llama-3.3-70b-versatile were deprecated for free and developer tiers on 2026-08-16 and remain only for enterprise customers on a committed-spend contract. "Contact Sales" is the permanent answer, so the carried-forward $0.05/$0.08 and $0.59/$0.79 were dropped this pass and the host rows show speed only
- **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
- **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
- **Nvidia Nemotron, Microsoft Phi** — not fetched this pass
- **xAI Grok 4.20** — still 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** — mostly closed this pass: per-model pages now state windows for GPT-5.5, GPT-5.4 mini, six Gemini rows, four Mistral rows, four Z.ai rows and the three Sonar rows. Still not stated: max output on the Mistral rows, Qwen3.8-Flash's and Qwen3.7-Plus's windows, and Devstral 2's, which is retired
- **Mistral — caching mechanics** — narrowed again. The caching doc now states 64-token cache blocks — a prompt under 64 tokens never hits — so min_tokens is 64 on every Mistral row. It still does not say whether caching applies without prompt_cache_key or give a TTL, so caps.caching.kind stays 'unverified'
- **Z.ai GLM-5.1** — still on the pricing page this pass at $1.4/$0.26/$4.4, unchanged, and still not added — the board carries a curated set per lab. GLM-5.3-FlashX now appears beside GLM-5.3-Flash at $0.37/$0.075/$1.25; not added for the same reason. Neither was ever a board row, so nothing on the board changes
- **Fields carried forward, not restated this pass** — two now. Qwen3.7-Max's 65,536 max output (no Alibaba page restates it, and the OpenRouter page fetched this pass returned API docs without the figure). And the three Sonar rows' context windows — 128K, 200K and 128K — which were stated on the Sonar model pages until 2026-09-27, when those pages became the Agent API migration guide; the Agent API models table has no context column
- **Audio input on the GPT-5.6 rows** — unchanged: the GPT-5.6 and GPT-6 model pages list "Input modalities: text, image" and no audio feature, which is still not a page saying there is none. The new GPT-6.1 Sol row carries audio_in as not stated, like the other GPT-6 rows
- **Listed on first-party pages this pass, not added** — Mistral's Ministral 3 (14B / 8B / 3B), Perplexity's Sonar Deep Research (Chat Completions, now mapped to the Agent API high preset), OpenAI's GPT-5.6 Cyber, GPT-Live-1 and gpt-rosalind-research (approved research access only, billing from 2026-10-05), Kimi K2.7 Code HighSpeed (noted on the K2.7 Code row), Z.ai GLM-5.3-FlashX ($0.37/$0.075/$1.25, still on the pricing page with no release note or date), Qwen3.8-2.4T-A95B on Together at Alibaba's own $2.00/$0.25/$6.00, and Qwen3.8-Omni-Flash (2026-09-18) and its realtime variant (2026-09-21), omni-modal models outside a text-token comparison. Mistral now also lists Z.ai GLM 5.3 as a third-party open-weight model it serves, and its hosted GLM 5.2 retires there on 2026-10-31; no Mistral rate for either was sourced, so Mistral is not a host row for them. The board carries a curated set per lab; recorded here so each omission reads as a choice
- **Gemini 3.8 Live, 3.8 Live Extended Thinking, and the 3.8 TTS models** — Live is unchanged since last pass: $0.75 text / $3.00 audio / $1.00 image-and-video input and $4.50 text / $12.00 audio output per 1M tokens, Live API audio-to-audio, outside this board's speech card. Gemini 3.8 Flash TTS and 3.8 Flash-Lite TTS went GA on 2026-09-22, free through 2026-12-31 and then $1.00 text in / $18.00

and $12.00 audio out. Speech synthesis is sold per output audio, not a board row. Recorded so the omissions read as choices

- **Xiaomi MiMo — candidate provider** — carried again. mimo.mi.com still prices MiMo-V2.6-Pro at $0.435 in / $0.0036 cache hit / $0.87 out with a 1M window, unchanged. The weights recorded here last pass need a closer name: Hugging Face has MiMo-V2.6-Pro-RL (2026-09-21) and MiMo-V2.6-Pro-MOPD (2026-09-27), both MIT, but no repo named MiMo-V2.6-Pro, so which weights the API serves is not stated. Adding a provider still means sourced provider notes, not one row; logged for the 2026-10-08 pass
- **Alibaba decision-model-preview** — a Jev-class model from Alibaba Cloud, released 2026-09-24 in Model Studio: classification, yes/no and scoring in parallel over text, returning probabilities and confidence scores, input-only billing, 65,536-token text input, 1,200 RPM. Its only price is "Limited-time free", with no end date and no list rate, so there is no number to put in a row; it becomes one when Alibaba publishes a rate. The typed-decisions card names it
- **Open-weight Jev-class models** — about twenty open-weight replications of the System One interface appeared on Hugging Face between 2026-09-16 and 2026-09-23 — LoRA adapters and fine-tunes on Qwen3.5, Gemma 4 and ModernBERT, nearly all Apache-2.0, several exposing a /v1/systemone endpoint (JevK5, Hopper, Winnow-12B, kev, Xor, Laya, Von and others). None has a public host and price, most are single-author, and six unrelated projects share the name OpenJev. Not rows until one does; recorded so the job card's 'open-weight options exist' is a sourced statement. Hosted resellers of Jev itself — Vercel AI Gateway, Netlify AI Gateway, classifier.dev — are noted on the Jev row rather than carried as hosts. Naderu's own naderu-laya-150m (bytesbrains/naderu-laya-150m, Apache-2.0, published 2026-09-28) is held to the same rule: it is a fixed five-way router fine-tuned from ModernBERT-base with no hosted price, not a general typed-question model, so it is disclosed in the band marked as ours below the news and is not a row

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

2026-10-22

- Weekly refresh: 69 rows re-verified and re-dated to 2026-10-22 against first-party pages fetched this pass; 8 rows carried at 2026-10-15 (see coverage.deferred). 8 rows added, 1 removed, 2 demoted: 77 rows total.
- Added Gemini 4 Argon at introductory $2.00/$0.10/$10.00 through 2026-11-30 ($4.00/$0.20/$20.00 from 2026-12-01), paid tier only, implicit caching exclusively, 1,048,576 unified window.
- Added Muse Wonder 1.0 (preview) at $4.00/$0.20/$20.00 standard-only, and Muse Spark Open (host-priced via Together at $1.25/$0.15/$4.25, no Meta endpoint). Meta's provider note no longer says the first-party API serves Spark alone.
- Added Mistral Large 4 (preview) at $1.36/$0.14/$4.18 and MiniMax M3.1 Flash (GA) at $0.15/$0.015/$0.60; both API-only until their weights land with readable licences.
- Jev 1.13 moves to previous; added Jev 1.14 at the same $0.042 with the 262K opt-in long-context header. Perplexity Decider v1.1 (no stated rate left) is replaced by v1.2 at $0.02 plus Flash v1.2 at $0.01; typed-decisions job card updated.
- Meta contributor limit now reads 150 RPM (row said 100); DeepSeek rows re-cite the 1017 cutover; meragpt source moves to the canonical decider-1 page.
- No lab changed a list price this week apart from scheduled introductory windows already on the board. Kimi K2.7 Code and K2.6 could not be re-read (pricing tables render client-side); Gemini 3.7/3.5 Flash left the pricing page; Llama 4 Scout, Devstral 2 and the Sonar Pro rows likewise — all carried, none repriced.
- Jobs re-read: only typed-decisions changes (new picks). All other cards stand on this week's rows.

2026-10-15

- Weekly refresh: all 70 model rows and 3 host rows re-verified and re-dated to 2026-10-15.
- Dual-search split & vocabulary: established separate jobs web-search (Sonar, Gemini 3.8 Flash, Grok 4.7) and rag-search (Claude Sonnet 5.5, Gemini 3.8 Flash, GPT-6 Luna), with web-search-native and long-context-retrieval capability constraints (naderu#380).
- Claude Haiku 4.5 reaches its earliest stated retirement date of 2026-10-15 with no 60-day shutdown notice published; continues answering as Legacy.
- Re-verified unchanged and re-dated: every price on every OpenAI, Anthropic, Gemini, Mistral, DeepSeek, Grok, Kimi, Z.ai, Qwen, Meta, MiniMax, TypeSafe, meraGPT, milliseconds.ai, Together and Hugging Face row. Apart from previously tracked adjustments, no lab changed a list price this week.

2026-10-08

- Added three rows. Claude Haiku 5.5 ($0.10/$0.01/$0.125/$0.50 up to 100K input tokens, $0.50/$0.05/$0.625/$2.50 above; 1M / 128K; caches from 512 tokens), released 2026-10-07. OpenAI's Decisions API on GPT-6 Luna ($0.10 per 1M input, no output, cache or write charges; beta from 2026-10-06), a separate row because /v1/decisions bills differently from the GPT-6 Luna row. Perplexity Decider v1.1 27B ($0.02 per 1M input, output free; 262,144-token input), whose weights Perplexity does not state.
- Claude Sonnet 5.5 cache read $0.20 → $0.10 (0.1x → 0.05x input), per the 2026-10-07 release note and the pricing page's multiplier section; the same page's model table still prints $0.20, recorded in the row note.
- Decider 1 (meraGPT) input $0.03 → $0.04 per 1M.
- Generation latest → previous: Claude Haiku 4.5, which its own model page now labels Legacy. Retirement still not sooner than 2026-10-15 with no notice given; its 200K / 64K now sourced from its model page, since the overview no longer lists it.
- Job cards & dual-search vocabulary (naderu#380): split web search from enterprise RAG — added web-search (Sonar, Gemini 3.8 Flash, Grok 4.7) covering live internet retrieval and per-call search fees, renamed and refocused rag-grounded-answers → rag-search for private vector DB retrieval and prompt cache economics (Claude Sonnet 5.5, Gemini 3.8 Flash, GPT-6 Luna) to align with Cruise lane routing, and added web-search-native and long-context-retrieval to constraints. chat-assistant swaps Claude Haiku 4.5 for Haiku 5.5. typed-decisions now names Perplexity Decider v1.1, Jev and OpenAI's Decisions API, with decision-machine-1 and Decider 1 in the text. extraction notes Haiku 5.5. The saver and balanced postures, and the Anthropic, OpenAI, Perplexity and meraGPT provider notes, are rewritten against this week's rows.
- Correction — the retired Perplexity Chat Completions request fee was described as $5–$14 per 1,000 requests on the former rag-grounded-answers card; corrected to $5–$12 on the web-search card to match Sonar's row note, reports, and the 2026-10-01 changelog.
- Ledger: Z.ai's GLM-5.3-FlashX ($0.37/$0.075/$1.25) recorded on the pricing page, not added. Kimi's pricing tables, which render client-side, were read from the pages' markdown versions this pass.
- Re-verified unchanged and re-dated: every other price on every OpenAI, Anthropic, Gemini, Mistral, DeepSeek, Grok, Kimi, Z.ai, Qwen, Meta, MiniMax, TypeSafe and milliseconds.ai row, and every licence read from Hugging Face. Apart from the Sonnet 5.5 cache read and Decider 1, no list price moved.

2026-10-01

- Added two rows. GPT-6.1 Sol ($2.00/$0.10/$2.50/$10.00, 1.05M / 128K, above 272K input $4.00/$0.20/$5.00/$15.00; cache read 0.05x; effort low to max, no none), released 2026-09-29. Claude Sonnet 5.5 ($2/$0.20/$2.50/$10, 1M / 128K, caches from 512 tokens), released 2026-09-28.
- Generation latest → previous: GPT-6 Sol (its page now says "See GPT-6.1 Sol for the newer Sol model"; no shutdown date) and Claude Sonnet 5 (Legacy on Anthropic's overview; retirement no sooner than 2027-06-30).
- Perplexity: Sonar Chat Completions support ended 2026-09-27. Sonar moves to its Agent API rate, $1.00/$1.00 → $0.25 in / $0.0625 cache read / $2.50 out, with search billed at $0.0025 per web-search call instead of $5–$12 per 1,000 requests; its caching kind moves none → unverified, since a cache-read rate is now printed but no mechanism is. Sonar Pro and Sonar Reasoning Pro move to deprecated: the Agent API has no model of either name, and the migration guide maps them to the fast and low presets. Their rows keep the Chat Completions rates the pricing page still lists. All three context windows are carried forward, not restated.
- Gemini 3.8 Flash and 3.7 Flash restated from their own model pages: 1,000,000 / 64,000 → 1,048,576 / 65,536. Google did not change anything; the board had carried the latest-model page's rounded figures, and now matches the other six Gemini rows.
- Job cards: Claude Sonnet 5.5 replaces Sonnet 5 on content-writing, agentic-coding and rag-grounded-answers; GPT-6.1 Sol replaces GPT-6 Sol on agentic-coding and code-review. computer-use keeps Sonnet 5 and Opus 5 and now says why for both 5.5 models. rag-grounded-answers restates Sonar at its Agent API pricing; typed-decisions names Alibaba's decision-model-preview. The balanced posture and the Anthropic, OpenAI and Perplexity provider notes are rewritten against this week's rows.
- Ledger: Perplexity Agent API token rates closed (now sourced from the Agent API models table). Added Alibaba's decision-model-preview (limited-time free, no rate yet). Xiaomi MiMo carried again, with the weights question narrowed. The open-weight Jev-class entry now discloses Naderu's own naderu-laya-150m and why it is not a row.
- Re-verified unchanged and re-dated: every other price on every OpenAI, Anthropic, Gemini, Mistral, DeepSeek, Grok, Kimi, Z.ai, Qwen, Meta, MiniMax, TypeSafe, milliseconds.ai and meraGPT row, and every licence read from Hugging Face. No lab changed a list price this week.

2026-09-24

- Added four rows: GPT-6 Sol ($2.00/$0.20/$2.50/$10.00) and GPT-6 Luna ($0.10/$0.01/$0.125/$0.50), both 1.05M / 128K with a 2x-input, 1.5x-output tier above 272K input; Claude Opus 5.5 ($4/$0.20/$5/$20, 1M / 128K, cache read 0.05x); Grok 4.7 ($2.00/$0.50/$6.00, 500K, identical to Grok 4.6).
- New job, typed-decisions (Jev-class), with three new rows and three new providers: TypeSafe's Jev 1.13 ($0.042 in, output free, 64K per request), milliseconds.ai's decision-machine-1 ($0.04, 20,000 characters) and meraGPT's Decider 1 ($0.03, 4,096 tokens). A deliberate vocabulary addition — the fifteenth job slug — made in check_board.py in the same change. Open-weight replications are logged in the deferred ledger, not carried as rows.

- Generation latest → previous: GPT-5.6 Sol, Terra and Luna (OpenAI's models page now recommends GPT-6 Astra, Sol and Luna; none of the 5.6 rows is deprecated), Claude Opus 5 (Legacy on Anthropic's overview), Grok 4.6.
- Gemini 2.5 Flash: Google limits the 2.5 models to existing users since 2026-09-18. Row keeps its generation and gains the restriction in best-for and watch-out. Gemini 3.1 Flash-Lite gains its 2027-05-07 earliest shutdown.
- Newly stated on first-party pages, filled from null: context windows on GPT-5.5 (1.05M/128K), GPT-5.4 mini (400K/128K), six Gemini rows (1,048,576/65,536 each, from new per-model pages), four Mistral rows (256K; Codestral 128K), four Z.ai rows, and the three Sonar rows; max output 128K on Claude Opus 4.8 and Sonnet 4.6; Kimi K3's cache write ($3.00, 5-minute TTL).
- Correction — GLM-5.2's max output moves 131,000 → 128,000. The 131K came from a secondary source; Z.ai's own GLM-5.2 guide says 128K. Z.ai did not change anything.
- Caps and licences re-read first-party: xAI, Z.ai, Meta Muse Spark and MiniMax caching is automatic (was explicit or unverified on 12 rows); Mistral states a 64-token cache block, now min_tokens on every Mistral row; Kimi K2.7 Code and K2.6 batch at 60% of standard; GLM-5.3-Flash is open, MIT, multimodal and always-reasoning; MiniMax M3 is open under the non-commercial MiniMax Community License; Kimi K3 carries its own Kimi K3 License, not Modified MIT; GLM-5.3's licence text is now readable; Qwen3.5-397B-A17B's caching claim is withdrawn to unverified, because Model Studio lists it in neither caching mode.
- Llama 4 Scout's 10M window is now sourced to Meta's own Llama 4 model card, closing the ledger entry that had been open for five passes. Together's Scout fine-tuning rate is now $2.10 SFT with a $6.00 minimum.
- Hosts: Groq's two Llama rows drop the $0.05/$0.08 and $0.59/$0.79 carried since 2026-08-06 — Groq's deprecations page says both are enterprise-only since 2026-08-16, so "Contact Sales" is the answer, not a gap. Groq Qwen3.8-27B context corrected from a mistyped 131,042 to 131,072. Together raised Qwen3.7-Max to $1.50/$0.30/$4.50 (from $1.25/$0.13/$3.75) and lists Qwen3.8 Flash at $0.09/$0.28, below Alibaba's own rate.
- Sources moved: xAI docs (/docs/models → /developers/pricing and per-model pages), Kimi's three pricing pages (merged into /docs/pricing/chat), Perplexity pricing (/getting-started → /docs/getting-started), Muse Spark (developer.meta.com is behind a login; now dev.meta.ai/docs/models.md). DeepSeek's V4-Pro continuation is re-cited to its changelog — the pricing page lists the rate but not the sentence. DeepSeek's peak note now says what the page says: Chinese public holidays are off-peak in full.
- Job cards: GPT-6 Luna replaces GPT-5.6 Luna on chat-assistant, content-writing, document-understanding, rag-grounded-answers and content-moderation; GPT-6 Sol replaces GPT-5.6 Terra on agentic-coding and code-review; Claude Opus 5.5 replaces Opus 5 on code-review. Summarization's catch drops the 'no window for GLM-4.7-FlashX' line, which is now 200K. All three postures and the long-context, open-weights, low-latency and batch-api constraints rewritten against this week's rows.
- Perplexity: Sonar Chat Completions still retires 2026-09-27, three days from this pass; token rates unchanged. The 2026-10-01 pass decides what the three Sonar rows become.
- Re-verified unchanged and re-dated: every price on every OpenAI, Anthropic, Gemini, Mistral, DeepSeek, Grok, Kimi, Z.ai, Qwen, Meta, MiniMax and Sonar row. No lab changed a list price this week.

2026-09-17

- CORRECTION — DeepSeek-V4-Pro is not retired. Last pass marked it deprecated on the strength of the 2026-09-10 release note, which said deepseek-v4-pro requests would route to V4.1-Flash from 04:00 UTC on 2026-09-14. DeepSeek's pricing page and changelog now say the V4-Pro API continues after that date with billing unchanged, "in response to user demand", until further notice. Generation moves deprecated → previous; the peak $1.32/$0.044/$3.96 rate is unchanged and is again the rate that name bills. The original release note still describes the cutover, so the two pages disagree — this board follows the pricing page. Provider notes, the V4.1-Flash watch-out, and the mitigation line are rewritten to match.
- deepseek-v4-flash stays deprecated: retired 2026-09-10, name still answers from V4.1-Flash at V4.1-Flash's rate. That half of last week's story is still true.
- Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking are stable on Google's models page (last updated 2026-09-15) and priced on the pricing page. Not added as board rows — Live API audio-to-audio, outside this board's speech card. Logged in news and the deferred ledger. gemini-3.1-flash-live-preview is now labelled legacy on the models page.
- Perplexity: Sonar Chat Completions still retires 2026-09-27 — ten days from this pass. Token rates unchanged. Row watch-outs and a news item now name the Agent API migration explicitly.
- Mistral: now resells GLM 5.3 alongside GLM 5.2 on its API at the same $1.4/$0.14/$4.4; provider note updated. Cached rates on Medium 3.5 / Small 4 / Codestral / Large 3 re-confirmed against docs.mistral.ai/inference/pricing.
- Hosts: Together lists DeepSeek V4.1 Flash at $0.30/$0.006/$1.20 — matching DeepSeek's own peak rate on in, out and cache — added as a host row. Groq: Qwen 3.6 27B is gone from console.groq.com/docs/models; Qwen3.8-27B appears in Preview at $0.80/$4.00 and replaces that host row.
- Re-verified unchanged and re-dated: every OpenAI, Anthropic, Gemini (text), Mistral, DeepSeek Flash, Grok, Kimi, Z.ai, Qwen, Muse Spark, MiniMax M3 and Sonar row on the board. Llama 4 Scout's 10M window still unstated on the cited secondary source (fifth pass); licence half re-confirmed.

2026-09-10

- Added DeepSeek-V4.1-Flash at $0.30/$0.006/$1.20 peak ($0.15/$0.003/$0.60 off-peak), 1M window and 384K max output, vision, and MIT weights stated on its own Hugging Face card. Released today under the new name deepseek-flash.
- DeepSeek-V4-Flash → deprecated: retired today. Its name still answers, served by V4.1-Flash and billed at V4.1-Flash's price, so the row's price moves $0.44/$0.014/$1.32 → $0.30/$0.006/$1.20 — the rate that name now bills. DeepSeek-V4-Pro → deprecated: from 04:00 UTC on 2026-09-14 its requests route to V4.1-Flash at V4.1-Flash's rate, until a V4.1-Pro ships. Its own price is unchanged until then. The extraction and summarization cards and the caching example move to V4.1-Flash.
- Added GPT-6 Astra at $10.00/$1.00/$12.50/$50.00 (long context $20.00/$2.00/$25.00/$75.00), 1.05M window, 128K max output, reasoning effort low to max with no none. OpenAI's models page now names it as the place to start. GPT-5.6 Sol stays latest: it is still listed among OpenAI's flagship models, and its price did not move.
- Added Gemini 3.5 Flash-Lite at $0.30/$0.03/$2.50, deferred last pass as a suspected misread of Gemini 2.5 Flash's row. It is not: the pricing page gives it its own section, and its one input rate covers audio, which 2.5 Flash and 3.1 Flash-Lite price separately. RESOLVED from the deferred ledger.
- Added Qwen3.8-Flash at $0.15/$0.47 from Model Studio's pricing page, deferred last pass for want of a citable rate. No window, cached rate or capabilities are stated, and the row says so. RESOLVED from the deferred ledger.
- Caching mechanics, sourced this pass where a row previously had none or had it wrong. GPT-5.6 Sol, Terra and Luna: automatic → both (OpenAI's caching guide describes implicit and explicit-only modes), with a 1,024-token minimum and a 30-minute TTL, all previously blank. Kimi K3, K2.7 Code and K2.6: explicit → automatic ("Supports automatic context caching" on each model's page). Qwen3.8-Max: explicit → both, and its $0.25 cached figure is now labelled as the implicit rate — the explicit read is $0.17 and the 10%/20% multipliers this row cited do not apply to the qwen3.8 line. Six Gemini rows gain the minimum prefix from Google's caching page: 4,096 on the 3.x Flash rows and 3.1 Pro Preview, 2,048 on 2.5 Flash.
- Capabilities read off first-party pages, all previously blank or unverified: Kimi K3 reasoning → always; K2.6 reasoning → controllable and vision → true; K2.7 Code vision → true. Grok 4.5 and 4.3 reasoning → controllable and vision → true, Grok Build 0.1 vision → true, and Grok 4.3 batch discount → 20% — all from the model data embedded in xAI's models page. MiniMax M3 vision → true.
- MiniMax M3 tools: true → not stated. OpenRouter's page says the endpoint it shows does not accept tools, and the row's price comes from a different host. Recorded in the ledger.
- GLM-5.3-Flash: the launch promotion ended on 2026-09-09 as scheduled and is gone from the pricing page. The row already showed the $0.15/$0.50 list rate; its note, the extraction card and the saver posture no longer describe a promotion.
- Corrections to the board's own text. Gemini 3.1 Pro Preview's price note said "promotional through 2026-12-31, per the pricing page" — the page's section for that model carries no such wording, and the phrase is removed. Qwen3.7-Max's "limited-time 50% off" label is gone from Alibaba's page, and the row stops calling its rate a promotion. GPT-5.4 mini's watch-out said GPT-5.6 Luna was dearer per token; it has been cheaper on both sides since Luna's 80% cut. Devstral 2's watch-out gave a 2026-06-30 retirement date that contradicted its own note; it now says 2026-07-31, matching Mistral's deprecation table. Groq's notes carried a stale 840 tok/s (now 560) and a 131K-everywhere cap that MiniMax M2.7's 196,608 breaks.
- The gpt-5.6 long-context threshold narrows rather than closes: 272K is on OpenAI's pricing page now, but only in the Cyber models table's column labels, on its copy of the Sol row.
- Hosts: Together's legacy 'DeepSeek V4 Pro' row at $1.74/$0.20/$3.48 is gone from its table, and Kimi K2.7 Code is off its serverless table — listed only for fine-tuning now — so that host row is removed. Together lists Qwen3.7-Plus again, at $0.32/$1.28. Groq tags both Llama rows Enterprise, still "Contact Sales".
- Job vocabulary, pass 3 of 3 (naderu#304) — all fourteen slugs now ship with a card: deep-reasoning, code-completion, code-review, speech, rag-grounded-answers, computer-use and content-moderation. code-completion has two picks, not three, because only two rows carry a first-party fill-in-the-middle statement; speech covers only audio input on token-priced rows and says so, since transcription and synthesis are sold by the minute.
- Two slugs renamed back to the published vocabulary: summarisation → summarization and copy-editing → content-writing. The 2026-09-03 pass shipped both as a copy-edit, and Cruise vendors the #304 words with a CHECK constraint behind them — so for a week the board and its one known consumer disagreed on two slugs. The labels keep their prose; the content-writing card is widened from rewriting to drafting to match what the slug means. check_board.py now pins all three axes to the published lists, so a slug can only change on purpose.
- Two constraint cards corrected against this pass's data: batch-api said every lab outside OpenAI, Anthropic, Google and Mistral carries no batch discount — Grok 4.3 now carries xAI's 20%; audio-required said stated audio input is always priced apart from text — Gemini 3.5 Flash-Lite prices it at the text rate.
- Re-verified unchanged and re-dated: every row on the board. Carried fields that no page restated this pass are flagged on their rows and listed in the ledger.

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
Board JSON · Privacy · Terms
© 2026 BytesBrains Pte. Ltd. All rights reserved.