

## THE MODEL PULSE

ISSUE 25 · WEEK 41 OF 2026

## THE BIG READ

# Every price that fell this week fell for a segment; key your router on prompt length and call type, not list price

## KEY SIGNALS THIS PERIOD

3

Models added to the tree

7

Vendors shipped frontier-class

4

Architectural patterns crossed multiple vendors

5

Benchmark moves

## AUTHOR

**Brian Letort**

BrianLetort.AI

## PUBLISHED

**October 10, 2026**

Issue 25 · Weekly read

## WEB

**[brianletort.ai/industry/model](https://brianletort.ai/industry/model)**

The Model Pulse archive

## THE BIG READ

# The thesis this issue defends.

Independent industry analysis. Compiled from public model cards, vendor blogs, and leaderboards. The body sets the read for the rest of the brief.

Every list price that fell this week fell for a segment, so if you own a model router, key it on prompt length and call type rather than on list price. Claude Haiku 5.5 is the example: a tenth of Haiku 4.5's price below 100K prompt tokens and half of it above, on a tokenizer that counts about 30% more tokens for the same text, so the step arrives 20-23% sooner (at roughly 77K tokens as the old tokenizer counted them). The full rate cards, including GPT-6.1 Sol's six-times Ultrafast tier, the input-only Decisions meter and Mistral's launch sale, with the independent scores at each effort setting, are in the architecture watch below and in the benchmark tables. Haiku 5.5 is the one real capability step in the cheap tier (Terminal-Bench 4.0 from zero on Haiku 4.5 to 32.8% on an independent run at max effort), which is why its price matters. Mistral Large 4, a 1.05T-parameter mixture-of-experts model (one that activates a few sub-networks per token; 52B active) at \$0.68 / \$2.09 on a 50% launch discount, is a mid-tier price story: its output price is above \$1 on sale and at list. The 'price war' reading is Dbggr's and Latent Space's; the reading here is segmentation by prompt length and call type, and the per-segment question is which segment the volume lands in.

The decision tier answers part of that. OpenAI's Decisions API beta at \$0.10 per million input tokens arrived as a late and high-priced entrant into a hosted tier that already sat at or below four cents: TypeSafe's Jev at \$0.042 since mid-September, Microsoft-Decision-1 on Foundry at \$0.042 within 72 hours of OpenAI's beta (a post-trained Qwen3.5-9B, per Microsoft's post, outside the graded set), Cloudflare's Clef-flash reported at \$0.038, and Perplexity's pplx-decider-v1.1-27b under Apache 2.0 with a provenance manifest, while LM Studio 0.4.26 added a local /v1/decisions endpoint. One launch-day ladder (CTOL Digital) also lists Perplexity's hosted decider at \$0.02 per million input tokens; that figure is not verified against Perplexity's pricing page and is not used as the floor here. eesel AI and CTOL published the 'OpenAI is no longer the cheapest' ladder on October 9 and are credited for it. The segment that will absorb volume at this price is the short classify, route, approve and flag call, which is also the lowest-HBM, lowest-output-token call per unit of work; Codex's new opt-in telemetry

comparing Guardian V2 approvals against a Decisions verdict is the first harness measuring whether the approval path can move there. Agent Techniques carries the harness; the open question for OpenAI is whether GA comes with a cut, which the AI Stack Weekly puts at 41%.

The third thing the week settled is that the open-weight label and the open-weight artifact have come apart. Mistral Large 4 and Reflection Beam were both announced as open-weight frontier models; neither can be downloaded (the particulars are in the architecture watch). Mistral's weights are due at the end of October and no licence has been named; its current on-prem terms are a bespoke self-deployment agreement, which is why the weekly bets the published licence will not be Apache 2.0 or MIT. Beam's weights are promised under Apache 2.0 'later this month' behind a waitlisted API with no published price. The tree's response is procedural: Mistral Large 4 is placed as a closed, active preview because anyone can call it, with its openness field to change when the repository exists; Beam is deferred, as MiMo-V2.6-Pro was in W39, until the weights can be checked against the card. The models that did ship downloadable weights this week were EmbeddingGemma 2 under Apache 2.0, Kandinsky 6.0 Video under MIT, Qwen-Image-2.1-Turbo under Qwen's research licence and the Perplexity decider; for a buyer planning on-prem deployment, the repository and its licence metadata are the deliverable, and the announcement is not.

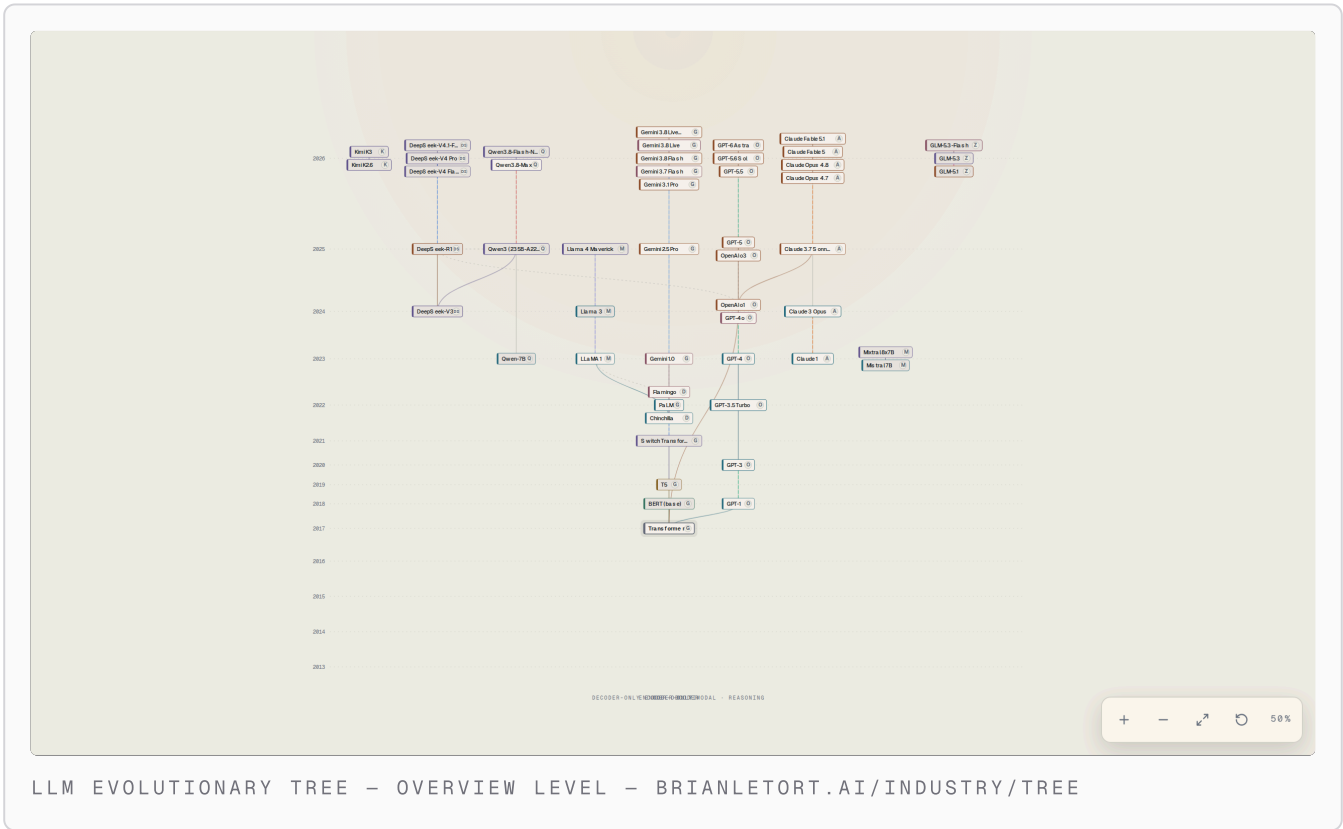
#### THE BAR

A model release matters when it changes one of three things: which workload runs on-prem, which tier of license is viable for a deployment, or which capability ceiling the closed frontier enforces. Single-vendor releases that change none of these are noise.

TREE DELTA

# What changed in the tree.

Three rows added (Claude Haiku 5.5 on reasoning under Haiku 4.5; Mistral Large 4 as a 1.05T / 52B-active multimodal MoE under Mistral Large 3, recorded as closed until its repository exists; EmbeddingGemma 2 on encoder\_only alongside text-embedding-3), one updated (Claude Sonnet 5.5, October 7 cache-read halving), four deferred and five excluded. The decision models still wait because the open condition in the placement rules' 3-of-6 test is architectural distinction, and Microsoft's disclosure that Decision-1 is a post-trained Qwen3.5-9B (per Microsoft's post, outside the graded set) supports the fine-tuning-recipe reading; the dispositions below carry the detail.



ADDED (3)

claude-haiku-5-5

mistral-large-4

embeddinggemma-2

UPDATED (1)

claude-sonnet-5-5

The decision-model category now has four confirmed hosted vendors (TypeSafe's Jev, OpenAI gpt-6-luna via the Decisions API, Microsoft-Decision-1, Cloudflare Clef and Clef-flash) plus one unverified (Perplexity's hosted decider, listed in a third-party ladder but not checked against Perplexity's pricing page), and at least two downloadable implementations (pplx-decider-v1.1-27b, the W40 llama.cpp task heads; Liquid AI's d1 pending verification); Microsoft's benchmark appendix (per its post, outside the graded set) names five more decision models not reviewed here (H2O-Lightning-4B, deck-31B, Strands-Decider 2B, Quyet-1.0-Large, Rune 26B-A4B). The independent-vendor condition of the 3-of-6 rule is met by any count. The remaining question is whether a typed-output classifier head on a shared base is an architecture or a fine-

*tuning recipe; Microsoft's Qwen3.5-9B disclosure is the first vendor statement on that question and points to recipe. The test will be applied in full when a vendor discloses a non-fine-tune build or a third decision model ships with a technical report.*

## FRONTIER MOVEMENTS

# Flagship-class releases.

3 releases this period. Vendor-stated frontier capability. The releases that reset the closed-source ceiling.

**2026-10-07**

RELEASE 01

Anthropic

FRONTIER

REASONING

## Claude Haiku 5.5

### Anthropic's first sub-dollar 5.5 model, with a two-tier rate card and a tokenizer that counts more

The question for a platform team is where its prompts fall relative to 100K tokens, because the price steps 5x there and the new tokenizer brings that step 20-23% sooner on the same text (around 77K tokens as the old tokenizer counted them). The rate card is in the architecture watch; Anthropic's blended estimate across its own customers' mix is about 75% cheaper (vendor figure), which is the number to test against your own token mix. Capability is the reason to care: Terminal-Bench 4.0 goes from 0.0% on Haiku 4.5 to 32.8% on Artificial Analysis's independent run at max effort (39.2% on Anthropic's harness), and the independent index at max effort is 43.4, above GPT-6 Luna (38). Safeguards are not uniform across the 5.5 line: Anthropic's launch page says Haiku 5.5's cyber safeguards are more restrictive than Haiku 4.5's but somewhat less restrictive than its recent models', while its biology safeguards match Sonnet 5, Sonnet 5.5 and Opus 5. The model drops temperature, top\_p, top\_k, budget\_tokens and prefill, has 1M context with 128K output, and carries a retirement floor of October 7, 2027. It is also the obvious subagent and reviewer model for agent harnesses, which is the Agent Techniques angle.

SOURCE: ANTHROPIC

**2026-10-06**

RELEASE 02

Mistral AI

FRONTIER

MOE

## Mistral Large 4

### A trillion-parameter European MoE, callable today, downloadable at the end of the month under a licence not yet named

Buyers evaluating a sovereign-compute option should treat this as a preview with two moving parts: a price that doubles when the launch discount ends (\$0.68 / \$2.09 to \$1.36 / \$4.18; Mistral has not published the end date) and a licence that has not been named, with on-prem access today under a bespoke self-deployment agreement. The model is 1.05T total / 52B active with a 1.6B vision encoder, 1M context per Mistral's docs (Artificial Analysis measures the preview at 524K) and 160+ languages, trained from scratch on 3,800 Grace Blackwell GPUs in Mistral's own European data centres, one of two frontier-scale training-footprint disclosures this week (Reflection's is the other). Independent scores set the level: 38 on the Artificial Analysis index and 22.73% on Vals AI's Terminal-Bench 4.0 against Mistral's own 28.3%. Artificial Analysis and Willison's 'Mistral is back, about six months behind' frame is theirs; the addition here is that at \$1.13 per index task (Artificial Analysis, list price) it is dearer than the open Chinese models it trails, so the case is residency and provenance, not price or capability.

SOURCE: MISTRAL AI

2026-10-09

RELEASE 03

Microsoft

SPECIALIST

## Microsoft-Decision-1

**A post-trained Qwen3.5-9B decision model on Foundry at \$0.042 per million input tokens, matching TypeSafe's Jev and 58% below OpenAI's beta**

This is the datapoint that confirms OpenAI's Decisions API as the most expensive offer in its category three days after launch (eesel AI and CTOL Digital published the comparison on October 9), and it is why OpenAI is under pressure to cut at GA rather than hold price (the AI Stack Weekly puts a cut to \$0.05 or below at 41%). Microsoft's post, outside the graded set, discloses more than most entrants: the base is a post-trained Qwen3.5-9B, the comparison covers 36 benchmarks and about 150,000 questions against six other decision models, and the vendor-measured P50 latency is about 35x faster than GPT-6 Sol; no independent run exists yet, so the model is deferred on the tree with the category. For architects the relevance is in the approve, flag and route call: with Jev and Microsoft at \$0.042, Cloudflare reported at \$0.038, Perplexity's open 27B at zero marginal licence cost and OpenAI at \$0.10, the price of a verdict sits at or below four cents regardless of which cloud the agent runs in. For the weekly's small-reviewer prediction (March 31, 2027), a named sub-10B decision model at a major agent-platform vendor is a candidate that is not yet documented in Copilot's approval path.

SOURCE: MICROSOFT FOUNDRY (COMMAND LINE BLOG)

## OPEN WEIGHTS

# Open-frontier and open-source drops.

4 releases this period. Open-weights drops that change procurement options. Pull these into pilot when score parity meets license parity.

2026-10-05

RELEASE 01

Reflection AI

OPEN FRONTIER

MOE

## Reflection Beam

**A 501B / 23B-active coding MoE pretrained in under four weeks, announced as open-weight with nothing yet to download**

Do not plan a deployment on this until the repository exists. The architecture and training disclosures are the substantive part: 23.8T tokens in under four weeks on 6,144 GB300 NVL72 GPUs, then 100M+ RL rollouts on 10.5K GB300s with about 1.3B sandboxes, which is a data point on how fast a frontier-scale pretrain now runs on a Blackwell Ultra fleet. Every score is vendor-reported, and on both agentic-coding rows its own table shares with the Chinese open models (DeepSWE v1.1, tabulated below, and Terminal Bench v2.1) Beam trails all three, which the-decoder's 'most capable open-weight model built outside China' frame does not reconcile. Access is a waitlisted API with no price; weights, model card and technical report are promised under Apache 2.0 after red-teaming. Deferred on the tree.

SOURCE: REFLECTION AI

2026-10-06

RELEASE 02

Google DeepMind

EDGE / SMALL

MULTIMODAL

## EmbeddingGemma 2

**A 740M Apache 2.0 embedder that puts text, code, image, video and audio in one 768-dimensional space and runs in 567 MB on a phone**

Teams running retrieval on-device or at the edge get a single embedding space across four modalities plus code at a size that fits a handset: 270M text backbone plus optional 170M vision and 300M audio encoders, 8,192-token context, Matryoshka truncation to 512, 256 or 128 dimensions, about 191 MB active RAM text-only and 567 MB fully multimodal on a Pixel 11 Pro. Google reports MT EB (Code) at 78.68 against 68.76 for the first EmbeddingGemma, a 14% gain, and 61.36 on multilingual MT EB; both are vendor-run, and the community's in-browser WebGPU demos within a day are evidence of loadability rather than quality. EmbeddingGemma 1 had more than 20M downloads, which is the distribution argument. Placed on encoder\_only, where the tree already carries text-embedding-3.

SOURCE: GOOGLE (THE KEYWORD)

**2026-10-07**

RELEASE 03

Perplexity

**SPECIALIST****DENSE**

## pplx-decider-v1.1-27b

**The open decision model gets a provenance manifest and a 61.56 Decision Index, up from 56.4, and stays deferred as a category question**

Last week's provenance problem (shards byte-identical to a community checkpoint) is answered: v1.1 ships with training code, a data manifest and a release manifest with file checksums under Apache 2.0, an independent header audit confirms the 26.09B-parameter layout, and OpenRouter listed it the same day. Perplexity reports 61.56 on its own Decision Index against 56.4 for v1 and a figure more than 3.5 points above TypeSafe's Jev; the index is Perplexity's, so treat it as a vendor benchmark. For buyers it is the only four-cent-class decision model whose weights, data lineage and licence can all be inspected, which matters more in an approval path than in a chat window. Deferred on the tree with the rest of the category.

SOURCE: HUGGING FACE

**2026-10-06**

RELEASE 04

Kandinsky Lab

**SPECIALIST****MULTIMODAL**

## Kandinsky 6.0 Video

**A 3B Lite and 29B Pro video-plus-audio family released under MIT with eight Diffusers checkpoints**

Outside the tree's scope but inside the open-weights pattern of the week: Kandinsky Lab shipped code, weights and Diffusers integration for a 5-second video generator with synchronized 44 kHz audio and lip-sync on the same days two Western labs announced language models that cannot be fetched. Eight checkpoints are on Hugging Face, including 10-step and 2-step distilled variants, and the README publishes per-GPU generation times (Pro at 1080p: 1,247 s on an RTX 4090, 402 s on an H100). For media teams the 3B Lite is the one that runs on a consumer card; no independent quality evaluation exists yet.

SOURCE: GITHUB (KANDINSKYLAB/KANDINSKY-6)

## ARCHITECTURE WATCH

# Patterns to track.

4 architectural patterns that crossed multiple vendors this period. Each pattern names the trend, the exemplar releases, and what it changes for deployment, cost, or capability.

**PATTERN 01 Price discrimination by prompt length, speed and call type**

Claude Haiku 5.5 (\$0.10 / \$0.50 to 100K prompt tokens, \$0.50 / \$2.50 above; cache \$0.01; tokenizer ~30% more tokens; Haiku 4.5 was \$1 / \$5 / \$0.10)

GPT-6 Luna (steps up at 272K input tokens: 2x input, 1.5x output)

GPT-6.1 Sol Ultrafast (\$12 / \$60 below 272K, \$24 / \$90 above; 6x Standard, 3x Fast)

Mistral Large 4 (\$0.68 / \$2.09 on a 50% launch discount with no published end date; \$1.36 / \$4.18 list)

OpenAI Decisions API (\$0.10 input, \$0 output, cache and reasoning)

This is the fact home for the week's rate cards. Five prices moved and none of them moved uniformly across a model's demand curve: Haiku 5.5 is a tenth of Haiku 4.5 for short prompts and half of it past 100K, where the newer tokenizer (which counts the same text as roughly 1.25-1.3x the tokens, per Anthropic's migration notes and Willison's measurement) arrives sooner; Luna steps up at 272K (2x input, 1.5x output); Sol Ultrafast sells the identical model's output at six times list for a vendor-claimed 8x speed; Mistral's price is a launch discount that reverts on a date Mistral has not published; and the Decisions API meters input only, which is attractive for long-document verdicts and irrelevant for short ones. The circulating 'price war reaches the bottom tier' frame is Dbggr's (October 10) and Latent Space's (October 8); the reading here is segmentation. The operational consequence is that a router keyed on list price will overpay on two of these five (Haiku above 100K, and Sol Ultrafast on batch work where nobody is waiting) and underprovision on a third (the decision tier, where the input-only meter makes long-document verdicts cheaper than list suggests). Key it on three fields: prompt-length band, latency class (Ultrafast only where a human is waiting), and call type (typed verdicts to the decision tier). On the prompt-length band, the operator's control is the compaction window: Zvi Mowshowitz's advice to set autocompact at 100K is the origin of 'compact before the step', and on the new tokenizer the step falls 20-23% sooner, at about 77K of old-tokenizer text; one practitioner report (wmedia.es, October 9) finds a 100K window thrashes in Claude Code and 130K is the working setting, so measure before pinning. Measure your own token mix under the new tokenizer before trusting any vendor blended-saving figure; Anthropic's own blended number for Haiku is 75% (vendor-reported), not 90%.

SOURCE: ANTHROPIC PRICING; OPENAI API CHANGELOG; MISTRAL AI

## PATTERN 02

## The decision-model category commoditizes before its largest entrant reaches GA

TypeSafe Jev (hosted, \$0.042 since mid-September)

OpenAI Decisions API on gpt-6-luna (\$0.10 input, beta Oct 6)

Microsoft-Decision-1 on Foundry (\$0.042, Oct 9; post-trained Qwen3.5-9B per Microsoft's post, outside the graded set)

Cloudflare Clef-flash (\$0.038, reported)

Perplexity pplx-decider-v1.1-27b (Apache 2.0, Decision Index 61.56)

LM Studio 0.4.26 local /v1/decisions endpoint

A category that did not have a name in August has, in six weeks, a typed interface (predicate, choice, score), four confirmed hosted vendors and one unverified (Perplexity's hosted price appears only in a third-party ladder), at least two downloadable implementations, a local-serving endpoint and a hosted price at or below four cents per million input tokens. The incumbent-then-clones sequence ran backwards: OpenAI's beta arrived late and priced highest. What is architecturally distinct is the output contract, not the network; every implementation whose base is disclosed is a classifier head post-trained on a shared 27B or smaller base (Microsoft's is a 9B Qwen, per its post), which is why the tree holds the category at the 3-of-6 test rather than cutting a branch. The infrastructure consequence is real regardless of placement: a verdict call emits a handful of tokens, needs little KV cache and no long reasoning trace, so volume that migrates here carries the lowest HBM and output-token intensity of any call type. If the approve/flag/block step of agent harnesses moves to this tier (Codex's Guardian-versus-Decisions telemetry is the first measurement of whether it can), aggregate request counts rise while tokens per request fall, which is the shape of demand that favours the accelerator vendor less and the network less than a chat-token boom would.

SOURCE: OPENAI API CHANGELOG; MICROSOFT FOUNDRY; PERPLEXITY; LM STUDIO

## PATTERN 03

## Open-weight label, closed artifact

Mistral Large 4 (API preview Oct 6; weights promised for the end of October, licence not yet named)

Reflection Beam (waitlisted API, no price; Apache 2.0 weights promised later in October; benchmark table revised Oct 8)

Contrast: MiMo-V2.6-Pro, Kolibri-1, EmbeddingGemma 2, pplx-decider-v1.1-27b (downloadable)

Two Western frontier-scale announcements used the open-weight label for models that exist only behind an API, while four smaller or non-Western releases in the same week shipped repositories with licence metadata; Marc Pope (October 7) and TensorFeed (October 6) made the cannot-download point first. The tree's handling is the useful precedent for procurement: a model is placed on what can be fetched, with the licence read from the repository rather than the press release, and a promised release date is a watchlist item, not a property. Mistral's technical documentation for downstream providers names no open-weights licence; it describes the current on-prem terms as a bespoke and confidential self-deployment agreement and says the document will be updated if weights are released, which is why the AI Stack Weekly bets that the published licence will be neither Apache 2.0 nor MIT. Reflection's situation is thinner: no price, no card, vendor benchmarks revised three days after launch. For a buyer the test is binary: can your platform team run `huggingface-cli download` today. For Mistral Large 4 and Beam the answer is no, and the open-frontier scorecard slots stay with models where it is yes.

SOURCE: MISTRAL AI TECHNICAL DOCUMENTATION FOR DOWNSTREAM PROVIDERS; REFLECTION AI

## PATTERN 04

## Trillion-scale totals, sub-5% active fractions, four-week pretrains

Mistral Large 4 (1.05T / 52B active, 5.0%; 3,800 Grace Blackwell GPUs)

Reflection Beam (501B / 23B active, 4.6%; 23.8T tokens in under four weeks on 6,144 GB300, vendor-reported)

MiMo-V2.6-Pro (1.02T / 42B active, 4.1%; W40)

Kolibri-1 (78.1B / 3.46B active, 4.4%; W40)

Four MoE releases in two weeks cluster at an active fraction between 4% and 5% regardless of total size, from 78B to 1.05T. Four points are a cluster, not a settled design point, but they are consistent enough to plan serving around: per-token compute looks like a 25-50B dense model while the weights need the memory of the full total, which is why these models are marketed with 748 GB desktop boxes and single-node B200 or H200 footprints rather than with FLOPS. The training-time disclosures are the newer datapoint. Beam's 23.8T tokens in under four weeks on 6,144 GB300 NVL72 GPUs is the first frontier-scale pretrain disclosed on a Blackwell Ultra fleet (vendor-reported, report pending) and the only one of the two with a wall-clock attached; Mistral's from-scratch run on 3,800 Grace Blackwell GPUs in its own data centres is the week's other footprint disclosure, without a duration. The capital consequence is in the AI Stack Weekly: compute that produces a frontier model in four weeks is financed years ahead on somebody else's balance sheet, and the HBM that the 4-5% active fraction leaves idle is paid for either way.

SOURCE: MISTRAL AI; REFLECTION AI

BENCHMARK MOVES

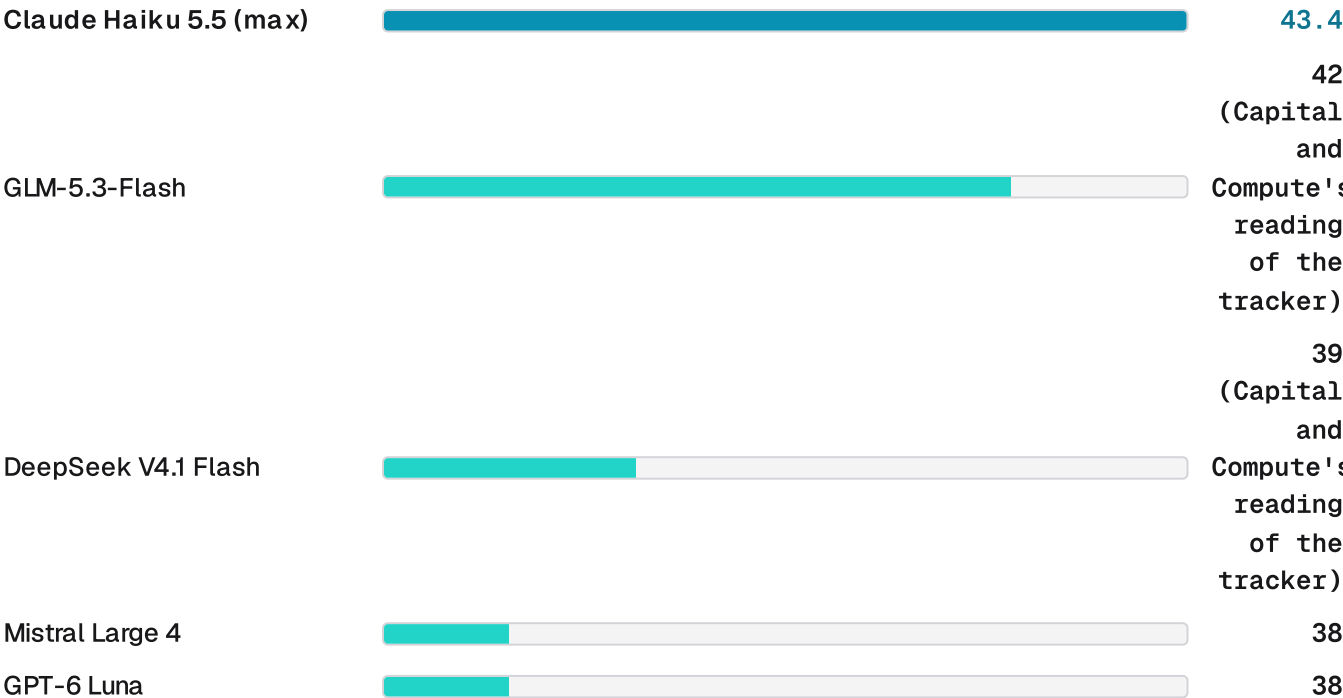
# Where the leaderboard moved.

5 benchmark deltas that change a procurement read. Scores reflect public leaderboards or vendor model cards as of publication.

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V4.3.2 (INDEPENDENT),  
CHEAP TIER

MOVE  
01

Haiku 5.5 at max effort enters the sub-dollar tier five points above GPT-6 Luna; Mistral Large 4 enters level with Luna and behind both open Chinese models that cost less to run. Setting in parentheses is the reasoning-effort level; scores are not comparable across settings

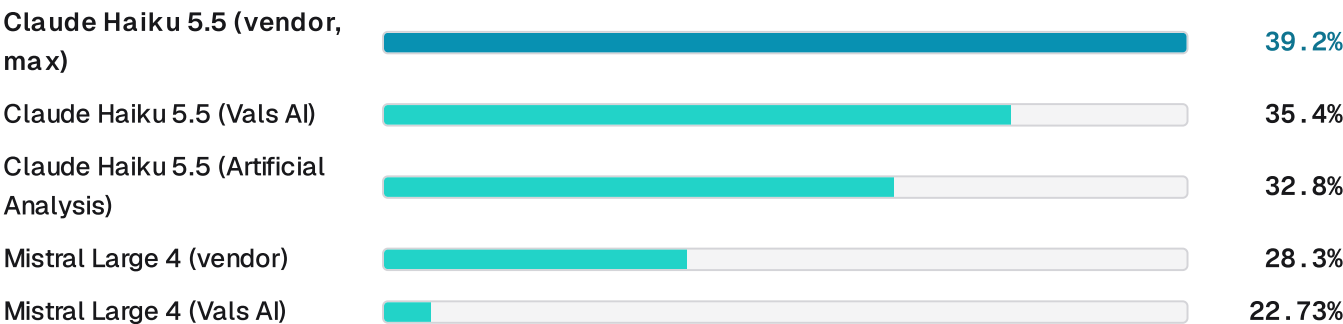


SOURCE: ARTIFICIAL ANALYSIS; GLM AND DEEPSEEK ROWS VIA CAPITAL AND COMPUTE

TERMINAL-BENCH 4.0, VENDOR HARNESS VERSUS INDEPENDENT RUNS

MOVE 02

Both new entrants score materially lower on independent runs than on their own harnesses; the gap is 6.4 points for Haiku 5.5 and 5.6 for Mistral Large 4. The mid-tier reference is Sonnet 5.5 at 70.6% on Anthropic's harness (W40)

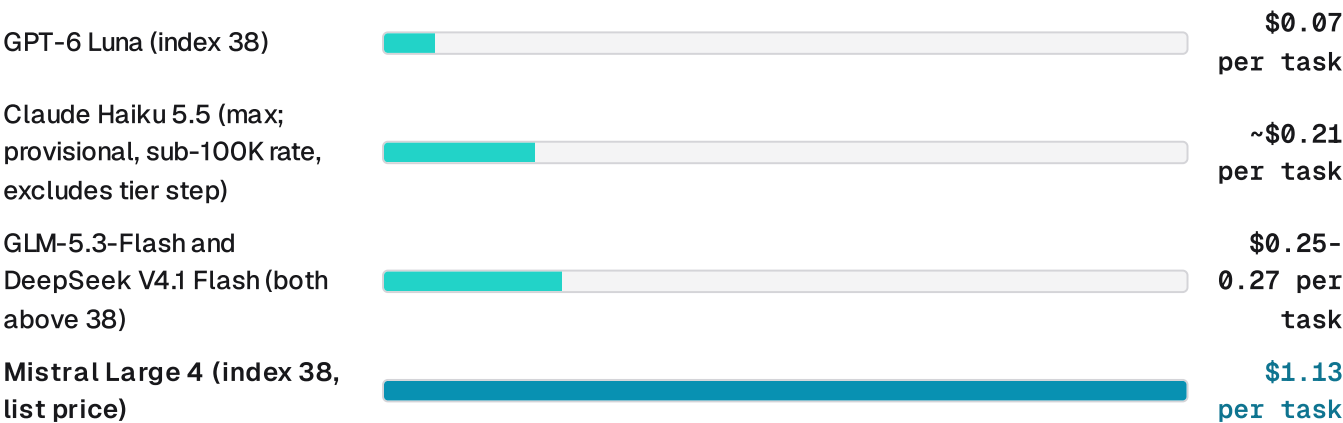


SOURCE: ANTHROPIC; MISTRAL AI; ARTIFICIAL ANALYSIS; VALS AI

ARTIFICIAL ANALYSIS COST TO RUN THE INTELLIGENCE INDEX (INDEPENDENT), CHEAP TIER

MOVE 03

Artificial Analysis's Haiku 5.5 cost figure is provisional and excludes the tier step: about \$0.21 per task at the sub-100K rate, three times Luna, as relayed by Capital and Compute from the tracker's Haiku 5.5 article, which also records Haiku at max effort using roughly 162K output tokens per task, about 3x Luna's; the row is carried as provisional rather than final and is outside the graded set. Mistral Large 4 costs about four times the two open Chinese models that outscore it

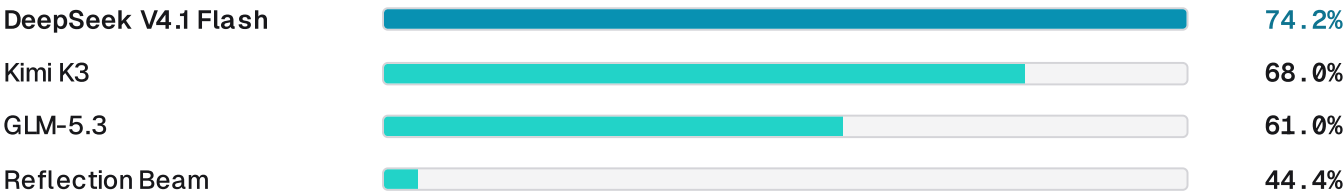


SOURCE: ARTIFICIAL ANALYSIS (HAIKU 5.5 ARTICLE, ARTIFICIALANALYSIS.AI/ARTICLES/CLAUDE-HAIKU-5-5) VIA CAPITAL AND COMPUTE

DEEPSWE V1.1, REFLECTION'S OWN TABLE (VENDOR-RUN, ALL ROWS)

MOVE 04

Reflection's October 8 revised table places Beam below the three open Chinese models it is marketed against on both agentic-coding rows shared across all four (DeepSWE v1.1 tabulated here; Terminal Bench v2.1 is the other); the figures are Reflection's and have not been reproduced

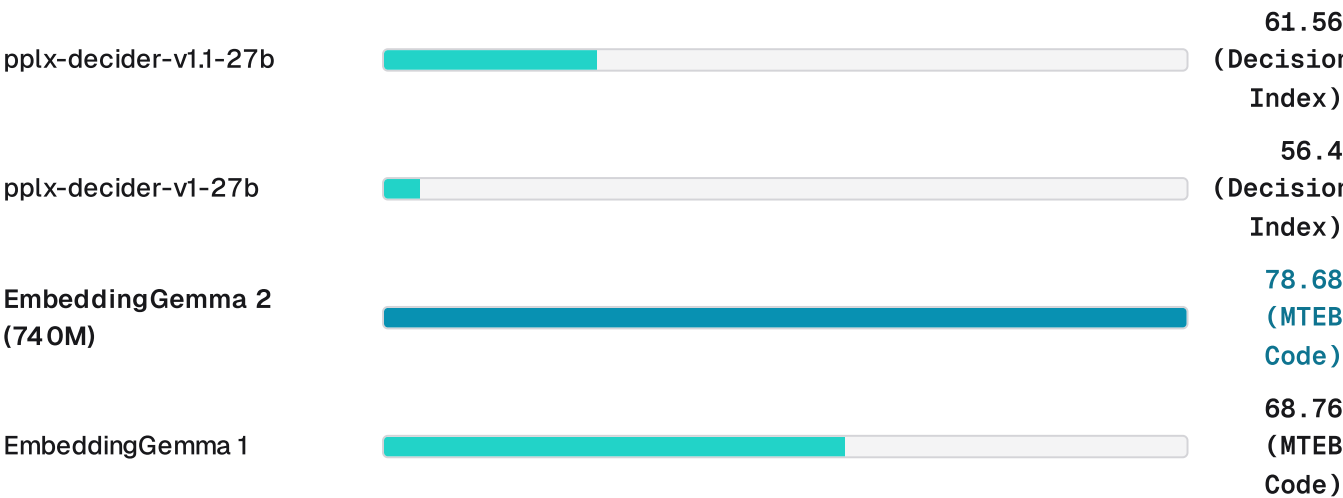


SOURCE: REFLECTION AI

PERPLEXITY DECISION INDEX (VENDOR-RUN) AND MTEB CODE (VENDOR-RUN)

MOVE 05

Two vendor-reported specialist moves: Perplexity's v1.1 decider gains 5.2 points on its own index, and EmbeddingGemma 2 gains 9.9 points on MTEB (Code) over its predecessor



SOURCE: PERPLEXITY; GOOGLE AI FOR DEVELOPERS MODEL CARD

TIER SCORECARD

# Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class.  
As of Oct 10, 2026.

| TIER                     | LEADER                                   | CHALLENGER        | READ                                                                                                                                                                                                                                                                                                        |
|--------------------------|------------------------------------------|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Closed frontier</b>   | <b>Claude Opus 5.5</b>                   | GPT-6 Astra       | Unchanged. Gemini 4 Argon ties Astra on the index but cannot be bought outside the Fairwind Program, so Astra holds the slot. No new top-tier model shipped; OpenAI's unreleased mathematics model has no API and GPT-6.1 Astra remains withheld. Independent index: Opus 5.5 at 58, Astra and Argon at 53. |
| <b>Open frontier</b>     | <b>MiMo-V2.6-Pro</b>                     | GLM-5.3           | Unchanged. Mistral Large 4 and Reflection Beam were announced as open-weight frontier models and neither is downloadable; the slots stay with models whose weights and licences can be checked. Mistral's independent index score (38) would in any case trail both.                                        |
| <b>Reasoning</b>         | <b>Claude Opus 5.5</b>                   | Claude Sonnet 5.5 | Unchanged. Sonnet 5.5's cache-read halving lowers its agentic cost by about 20% on Anthropic's estimate without changing its score.                                                                                                                                                                         |
| <b>Coding</b>            | <b>Claude Sonnet 5.5</b>                 | GPT-6.1 Sol       | Unchanged. Haiku 5.5's Terminal-Bench 4.0 (32.8% independent) is a cheap-tier result, not a challenge to the mid tier; GPT-6.1 Sol became the Codex default and gained an Ultrafast tier, which changes speed and price, not capability.                                                                    |
| <b>Multimodal</b>        | <b>Gemini 3.8 Live Extended Thinking</b> | Gemini 4 Argon    | Unchanged. Mistral Large 4 adds image input to a trillion-parameter MoE; its multimodal scores are vendor-only. EmbeddingGemma 2 embeds text, code, image, video and audio in one space and is tracked on the edge line rather than here.                                                                   |
| <b>Sub-dollar hosted</b> | <b>Claude Haiku 5.5</b>                  | GPT-6 Luna        | New tier, defined as hosted models with list output price below \$1 per million tokens at the base band. Haiku 5.5 leads on the independent index (43.4 at max effort against 38) and on independent Terminal-Bench 4.0; Luna leads on measured cost per index task                                         |

| TIER         | LEADER                  | CHALLENGER                                         | READ                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------|-------------------------|----------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              |                         |                                                    | (\$0.07 against a provisional ~\$0.21) and has the simpler rate card. GLM-5.3-Flash (42) outscores Luna but is tracked on the open-frontier line. Mistral Large 4 does not qualify at any price (\$2.09 output on sale, \$4.18 at list) and is tracked in the weekly's mid-tier lever.                                                                                                                       |
| Edge / small | EmbeddingGemma 2 (740M) | 4B decision heads via llama.cpp (Kev-4B and peers) | Leader changes. The tier covers models meant to run on one device or one CPU; EmbeddingGemma 2 takes the slot on an Apache 2.0 licence, a 567 MB on-device footprint demonstrated on a Pixel 11 Pro, and day-one community WebGPU builds. Its MT EB figures are vendor-run. The 4B decision heads move to challenger; Clef-flash (9B) drops off as a hosted price cut rather than an edge release this week. |

## VENDOR SIGNALS

# Pricing, gating, deprecation.

6 non-release moves that shift vendor risk — pricing, deprecations, gating decisions, license changes — each with a one-line procurement read.

**2026-10-07**

SIGNAL 01

Anthropic

**Halves Sonnet 5.5 cache-read price to \$0.10 per million (input, output and cache-write unchanged), citing a ~20% saving on most agentic tasks; adds monthly Claude Platform API credits of \$100-200 for Max and up to \$500 pooled for Team**

For agent operators this is the week's only unambiguous cut, because cache reads are the meter long sessions hit hardest and the cut applies to the model most agent harnesses already default to. The API credits are a bundling signal: Anthropic is routing subscription revenue toward platform usage, which blurs the seat-versus-token boundary that The Application Layer tracks.

SOURCE: ANTHROPIC

**2026-10-07**

SIGNAL 02

Anthropic

**Haiku 5.5 ships with breaking changes against Haiku 4.5: temperature, top\_p, top\_k, budget\_tokens and prefill removed; the same text counts ~30% more tokens; retirement not before October 7, 2027**

Migrations from Haiku 4.5 are not drop-in. Teams should budget a re-evaluation of any prompt that used sampling controls or prefill, re-measure token counts on their own corpus before accepting the 90% headline, and record the one-year retirement floor in vendor scorecards as a deployment guarantee.

SOURCE: CLAUDE PLATFORM DOCS

**2026-10-06**

SIGNAL 03

OpenAI

**Decisions API enters public beta on gpt-6-luna at \$0.10 per million input tokens with no output, cache or reasoning charge, up to 10x faster than Luna via Responses (vendor claim); GA 'in the coming weeks'; API usage tiers collapse from five to three**

Architects can now put a typed verdict call in a pipeline without running a classifier themselves, but the price is the category ceiling, not its floor, and the beta terms can change at GA. The tier simplification (Build \$5, Launch \$100, Grow \$500) lowers the friction for small accounts and is the kind of change that precedes a volume push.

SOURCE: OPENAI API CHANGELOG

**2026-10-08**

SIGNAL 04

OpenAI

**GPT-6.1 Sol Ultrafast reaches GA for all API users at \$12 / \$0.60 cached / \$60 per million below 272K input tokens (\$24 / \$90 above), with US and EU data residency; also in Codex and ChatGPT Work for Pro \$500 and eligible Enterprise and Edu plans**

Speed is now a purchasable SKU on the mid tier at a 6x premium. Operators should confine it to interactive paths where a human is waiting and keep batch and background agent work on Standard; the data-residency option makes it the first Ultrafast tier a regulated EU buyer can use.

SOURCE: OPENAI RELEASE NOTES

**2026-10-06**

SIGNAL 05

Mistral AI

**Mistral Large 4 preview priced at a 50% launch discount (\$0.68 / \$0.07 cached / \$2.09) on a \$1.36 / \$0.14 / \$4.18 list, with no published end date (a two-week duration circulates in secondary coverage of Mistral's changelog and is not in the graded set); weights promised for the end of October after red-teaming with cybersecurity partners and state authorities; no licence named**

Budget the list price, not the sale. The red-teaming-with-state-authorities language and the bespoke self-deployment terms that govern on-prem access today are the sovereignty positioning made concrete, and they are the terms a European buyer should read before the weights land; the AI Stack Weekly bets that the published licence is neither Apache 2.0 nor MIT.

SOURCE: MISTRAL AI

**2026-10-08**

SIGNAL 06

Local serving stack (vLLM, llama.cpp, LM Studio, Ollama)

**vLLM v0.31.0 (717 commits) with breaking changes to multimodal kwargs and FP8 quantization flags; llama.cpp day-zero builds for the week's open releases; LM Studio 0.4.26 adds a local /v1/decisions-compatible endpoint; Ollama 0.40.2**

Platform teams that pin vLLM should read the v0.31.0 migration notes before upgrading; the FP8 flag rename and the multimodal kwargs gate will break existing launch scripts. LM Studio's decisions endpoint means the four-cent verdict call can also be a zero-cent local one for on-device approval paths.

SOURCE: GITHUB (VLLM-PROJECT/VLLM RELEASES)

## WATCHLIST

# On the radar next.

6 model-side catalysts in the next 7–30 days that would change the read materially. Watching these tells us whether the canopy is widening or thinning.

**OPEN (MISTRAL HAS PUBLISHED NO END DATE)**

WATCH 01

**Mistral Large 4 launch discount ends; price reverts to \$1.36 / \$4.18**

The first test of whether the preview volume survives a 2x step; Mistral Studio usage and OpenRouter share are the public signals. A two-week duration circulates in secondary coverage and is not relied on here.

**OCT 27–31**

WATCH 02

**Mistral Large 4 weights and licence on Hugging Face; Reflection Beam weights, model card and technical report**

Two deferred or provisional tree rows resolve on repositories: Mistral's openness field flips to open\_weights if the files land, and the licence decides the AI Stack Weekly's bet that it is neither Apache 2.0 nor MIT (resolves November 15); Beam is placed or stays deferred on whether a card and checksums exist.

**Q4 2026**

WATCH 03

**Artificial Analysis publishes a tiered cost-per-task figure for Claude Haiku 5.5**

The tracker's current figure is provisional and excludes the step above 100K tokens. When the tiered number lands, it settles whether Haiku's effective cost in agent loops is nearer Luna's \$0.07 or the mid tier, which decides the sub-dollar scorecard row.

**COMING WEEKS (PER OPENAI)**

WATCH 04

**Decisions API general availability and its price**

A GA price at or below \$0.05 is the AI Stack Weekly's bet (by January 31, 2027); a hold at \$0.10 would mean OpenAI is selling the integration rather than the verdict, with the hosted tier already at or below four cents at TypeSafe, Microsoft and Cloudflare.

**Q4 2026**

WATCH 05

**Gemini 4 Argon developer availability outside the Fairwind Program**

Carried from W40. The scorecard's closed-frontier challenger line still carries a model nobody outside the programme can call; the AI Stack Weekly's open prediction on Argon availability resolves on the Gemini API model list.

OPEN

WATCH 06

## Independent Terminal-Bench 4.0 and index runs for Reflection Beam; withdrawal count on OpenAI's 560 unverified mathematics manuscripts

Beam's vendor table was revised once already and no third party has run it; the OpenAI corpus's error rate, not its size, is what would make the unreleased model a tree entry when it ships.

---

METHODOLOGY AND AUTHORSHIP

# How this brief is built.

Compiled from public model cards, vendor blogs, leaderboards, and official lab announcements. The publication is anchored to the LLM Evolutionary Tree at [brianletort.ai/industry/tree](https://brianletort.ai/industry/tree), which the brief annotates each period.

WHAT EACH SECTION IS FOR

|                           |                                                                                                                                                 |
|---------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| TREE DELTA                | Model rows added or updated in <code>content/llm-tree/models.yaml</code> since the prior issue. Every id is real and clickable on the web view. |
| FRONTIER AND OPEN WEIGHTS | Releases that reset the closed ceiling or move the open-frontier line. Each card cites a single primary source.                                 |
| ARCHITECTURE WATCH        | Patterns that crossed multiple vendors in the period. Named once, exemplified by recent releases.                                               |
| BENCHMARK MOVES           | Public leaderboards and vendor model cards. Bars reflect the score range across the rows shown, not zero-baselined.                             |
| TIER SCORECARD            | Leader vs challenger by tier. A snapshot for procurement shortlists; refreshed every issue.                                                     |

AUTHORSHIP

Brian Letort

BrianLetort.AI · Independent analysis. All sources public. Not investment guidance.

THIS ISSUE

Issue 25

Week 41 of 2026 · October 10, 2026

WEB

[brianletort.ai/industry/model](https://brianletort.ai/industry/model)

The Model Pulse archive, the LLM Evolutionary Tree, and the AI Stack Weekly companion publication.