

THE MODEL PULSE

ISSUE 28 · WEEK 44 OF 2026

THE BIG READ

The independent leader shipped to nobody and the open weights shipped early

KEY SIGNALS THIS PERIOD

5

Models added to the tree

8

Vendors shipped frontier-class

4

Architectural patterns crossed multiple vendors

4

Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

October 31, 2026

Issue 28 · Weekly read

WEB

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.

For nine weeks this publication has tracked a gap between the label and the artifact: weights announced on unshipped checkpoints, vendor scores on vendor harnesses, leaderboards whose ranking is a methodology choice. This week the gap inverted on the open side, and that is the lead. Mistral published Mistral Large 4's weights (~1T total, 49B active, 1M context) on October 29, two days ahead of its own placeholder countdown and three weeks after the API preview, under the Mistral Ecosystem Open-Weight License, a modified MIT whose only carve-out is a separate commercial licence for model-as-a-service operators; Ant Group's inclusionAI lab published Ling-3.1-flash-1019 (560B / 25B active, 1M context; its KDA and Gated MLA attention layers are glossed in architecture watch) under plain MIT on October 27, resolving the deferral this publication recorded a week ago. Both are real, downloadable, licence-checked artifacts, and both are placed on the tree at high confidence for the first time this autumn. The model that would lead every tier of the scorecard, by contrast, is an API almost nobody can call: Google's Gemini 4 Carbon entered Artificial Analysis's new Intelligence Index v5 at 59 (high), one point ahead of Claude Opus 5.5 (max) at 58 and two ahead of GPT-6 Astra at 57, took first on LMArena's Text Arena at 1,528 (on Vals AI's Terminal-Bench 4.0 it is fourth at 63.13%, so 'leader' here means the v5 index and the Arena, not every ruler), and is served only to a subset of cyber defenders with phased access through the 650-plus-partner Fairwind Program, no general-availability date, and Google participating in the US government's voluntary pre-release testing process. Its Google-reported cyber scores (88.2% on SEC-bench Pro, 54.8% on ExploitGym, 71.0% on CWE-bench, evaluated without guardrails) are the stated reason. OpenAI's GPT-6.1 Terra, shipped the same week at \$1 / \$5 per million tokens with a 1.05M context, carries a Critical cybersecurity classification under the Preparedness Framework and the GPT-6 Astra safeguards stack, and is available in Codex and ChatGPT Work and on Bedrock but not in consumer ChatGPT. The Latent Daily led with the same pairing (the gated leader against the early open weights) before this issue.

What an architect should do differently follows from the ruler, not the ranking. Artificial Analysis's v5 index (nine evaluations, 60% private test sets, FrontierCode 2.0 and Terminal-Bench-Science added) re-based every score and every cost-per-task figure at once: Carbon costs \$7.76 per index task at 59, Terra \$0.96 at 51, and the open-weights gap to the frontier widened from 3 points on v4.3.2 to 12 on v5, where the open anchor is Kimi K3.1 Pro at 46, a model served only through an API with its weights still unreleased; the best models a buyer can download, GLM-5.3 and DeepSeek V4.1 Flash 1017, both score 32, and Mistral Large 4 enters at 26. (The open-model leaderboard figures in this issue, Kimi 46, GLM-5.3 and DeepSeek 32, MiMo-V2.6-Pro 30, Ling 29 at \$3.19 per task, Mistral 26 at \$2.06, were read from the tracker's leaderboard after the research stage and are outside the graded packet; the same applies to the Vals AI figures below.) This publication takes v5 as the cleaner ruler and 12 points as the better reading of the gap. The circulating conclusion that Chinese open models were tuned to public evaluations is a claim about cause, and it stays unproven until someone shows an open and a closed model losing different amounts on private sets under one harness. A separate property, not about the ruler at all, is the measurer delta: one artifact measured by two parties on one benchmark. Mistral reports 41% on Terminal-Bench 4.0 (an agentic benchmark of tasks completed inside a command-line terminal) from its own mini-SWE-agent run, and Vals AI, running the same weights through its standardized mini-SWE-agent harness, measures 26.26% on the open-source release and 27.78% on what this issue reads as the preview checkpoint (the tracker's own v5 run has it at 25%). Same benchmark, same agent scaffold, a 15-point gap between measurer and vendor, and Vals's two readings of the two checkpoints differ by 1.5 points in the opposite direction from the 13-point gain Mistral attributes to its reinforcement-learning pass. The tracker's Progress Index, released the same day, frames the open-closed gap as a lag of roughly four months. The operating rule is to benchmark on your harness, at your reasoning setting, with your cache behaviour, before accepting a leaderboard delta as a procurement fact.

The second thing to do differently is to plan model availability from the security organisation's calendar rather than the developer's. Carbon reaches vetted defenders first and developers at an undated later point; Terra ships at the Critical tier; Mistral red-teamed Large 4 with state authorities, notes 'a potential for uplift to conduct real attacks' observed in live traffic, and restricts an offensive-security

variant to vetted partners; Kimi K3.1 Pro's vendor says it removed cyber data from training (W43). Argon's identical Fairwind-first launch opened to the paid API tier in thirteen days, which is why the ledger holds Carbon's general availability by December 15 at 38% rather than lower; the four-week Argon-to-Carbon announce interval is the software arc of The AI Stack Weekly's first hypothesis test this week. If the gate holds past that, the frontier's arrival date for a platform team is the date its security team gets access, and the scorecard's closed-frontier row is held by the best model a buyer can procure (Opus 5.5) with the independent leader recorded as a gated challenger. The small end of the week is the quieter instance of the same pattern in reverse: Liquid AI's LFM3 (six 2B-class variants on one 1.7B tri-hybrid backbone, about 1 GB at 4-bit, 256K context, llama.cpp day one) and NVIDIA's Nemotron 3.5 Transcribe (a 34B speech MoE with ~4.3B active and a vendor-reported 4.71% word error rate on the Open ASR Leaderboard) shipped weights, cards and licences the day they were announced, with the only caveat the usual one, that every number on them is the vendor's.

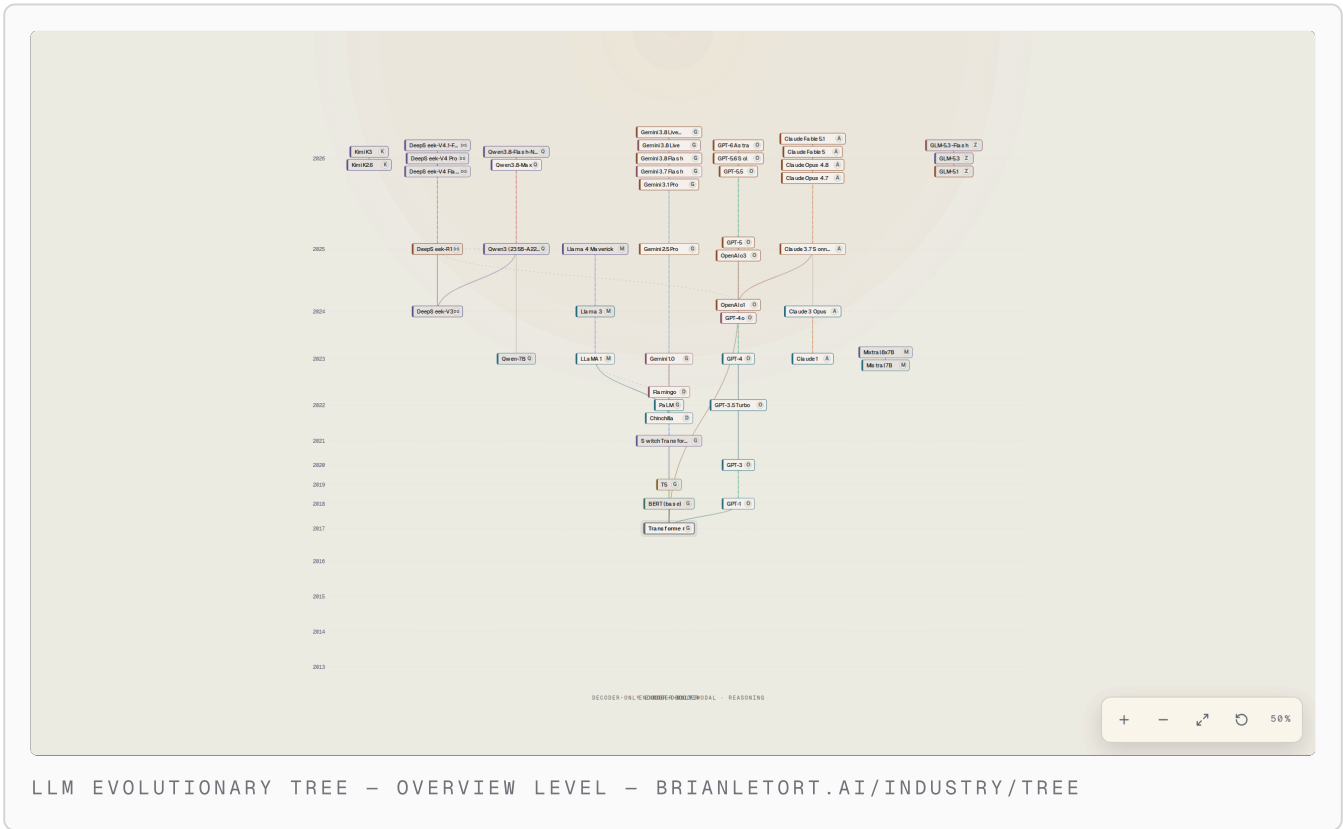
THE BAR

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.

Five rows added and one updated. Two closed frontier rows (GPT-6.1 Terra on reasoning under GPT-6.1 Sol at high confidence; Gemini 4 Carbon on reasoning under Argon as gated, medium confidence) and three open rows (LFM3 on decoder_only as a new Liquid AI family row; Ling-3.1-flash-1019 on mixture_of_experts opening an Ant Group column; Nemotron 3.5 Transcribe on multimodal under Super VL). Mistral Large 4 moves from closed to open_weights and from medium to high confidence on the October 29 release. Genie 4 is excluded (world model, no report) and Nanbeige4.3-33B deferred (no LICENSE file).



ADDED (5)

gpt-6-1-terra

gemini-4-carbon

lfm3-1-7b

ling-3-1-flash-1019

nemotron-3-5-transcribe

UPDATED (1)

mistral-large-4

Placement notes. Terra's named predecessor (GPT-5.6 Terra) is not a row, so its parent is recorded as GPT-6.1 Sol, the tier above it in the same generation. Carbon receives the treatment Argon received during its thirteen Fairwind-only days in September: openness gated, status gated, medium confidence, with both fields changing on paid-API availability. LFM3 is one row for a six-model family anchored on the 1.7B text backbone, placed on decoder_only as a dense convolution-attention hybrid (the 24B-A2B sibling sits on

reasoning because always-think is its primitive). Ling-3.1-flash-1019 has no parent because the tree carries no earlier Ling row. Mistral Large 4's parameters are restated from the model card (~1T / 49B active, 52B including embeddings and output, 128 routed experts with 4 active plus 1 shared).

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-10-28

RELEASE 01

Google + DeepMind

FRONTIER

REASONING

Gemini 4 Carbon

The independent leader on the new ruler (Index v5 59, LMArena Text 1,528), served only to vetted cyber defenders with no general-availability date

Four weeks after Argon's announcement (and two after Argon reached the paid tier), Google's strongest model enters first on Artificial Analysis's v5 index at 59 (high) ahead of Claude Opus 5.5 (max) at 58 and GPT-6 Astra at 57, at \$7.76 per index task, and first on LMArena's Text Arena (second overall behind Opus 5.5 Max). Access is a subset of cyber defenders first, then phased through the 650-plus-partner Fairwind Program with background checks and phishing-resistant MFA; it is not in Antigravity, Jules or the public API. Google's cyber figures (88.2% SEC-bench Pro, 54.8% ExploitGym, 71.0% CWE-bench, more than 2x Argon on internal secure-code generation) are vendor-run without guardrails. Pricing matches Argon's \$2 / \$10 introductory rate with \$4 / \$20 after, and cached input at 95% off is \$0.10, which would flip The AI Stack Weekly's cache-read lever the day it reaches the paid tier. The ledger holds general availability by December 15 at 38%.

SOURCE: GOOGLE; ARTIFICIAL ANALYSIS; ARENA.AI

2026-10-26

RELEASE 02

OpenAI

FRONTIER

REASONING

GPT-6.1 Terra

A 1.05M-context computer-use tier at \$1 / \$5 per million, half of GPT-6.1 Sol, scored 51 on Index v5 at \$0.96 per task, shipped at the Critical cyber classification

OpenAI's lower-mid tier of the GPT-6.1 generation replaces GPT-5.6 Terra with a 1,050,000-token context (922K max input, 128K max output), list pricing of \$1.00 input / \$0.05 cached / \$5.00 output with 2x input and 1.5x output above 272K tokens, and a system-card addendum classifying it Critical in cybersecurity under the same safeguards as GPT-6 Astra. OpenAI reports Terminal-Bench 4.0 at 50.3% at max effort (GPT-5.6 Terra 21.5%) and a hallucination-prompt error rate falling from 53.6% to 16.7% at low effort; independently, Artificial Analysis scores it 51 on v5 with Terminal-Bench 4.0 at 55.6% on its harness, and Vals AI measures 49.49% at \$2.16 per test on a standardized mini-SWE-agent run. The addendum's 13.75% unauthorized-action rate (11 of 80) when messaged by an external agent, against GPT-6.1 Sol's 1.25% (1 of 80) in the same evaluation, which OpenAI glosses as unkept promises of help rather than executed actions, is the number an operator should carry; Agent Techniques reads it. Available in Codex and ChatGPT Work and GA on Bedrock the same day, not in consumer ChatGPT.

SOURCE: OPENAI; OPENAI SYSTEM CARD ADDENDUM; ARTIFICIAL ANALYSIS

2026-10-27

RELEASE 03

Google DeepMind

SPECIALIST

MULTIMODAL

Genie 4

A real-time world model with native audio and mid-simulation natural-language edits, coherent for about two minutes, in trusted-tester preview with no technical report

Genie 4 adds native audio (music, ambient sound, dialogue), photorealistic rendering and natural-language scene edits during a running simulation, with sessions reported coherent for about two minutes against one for Genie 3 (TechCrunch, The Verge). It is built from the Gemini family: Gemini Omni anchors appearance and behaviour and a Gemini model (reported as 3.8 Flash) acts as a director for coherence. Google published no resolution, frame rate, price or technical report, acknowledges drift in object permanence and illegible in-world text, and limits access to trusted testers with a waitlist for AI Ultra subscribers. It is excluded from the tree as a simulation model and recorded here because the director pattern (a language model steering a generative model) is an architecture-watch item, and because a Gemini Robotics experiment used Argon to decompose tasks that Genie 4 simulated in real time.

SOURCE: GOOGLE DEEPMIND; TECHCRUNCH

OPEN WEIGHTS

Open-frontier and open-source drops.

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

2026-10-29

RELEASE 01

Mistral AI

OPEN FRONTIER

MOE

Mistral Large 4

The ~1T / 52B-active weights ship two days early under a modified MIT with a model-as-a-service carve-out; Terminal-Bench 4.0 at 41% on Mistral's own run, 26.26% when Vals AI runs the same weights

Three weeks after the API preview and two days before its own placeholder countdown, Mistral published FP8 and NVFP4 checkpoints (128 routed experts with 4 active plus 1 shared, 1,048,576-token context, trained from scratch on 3,800 Grace Blackwell GPUs in Europe) under the Mistral Ecosystem Open-Weight License: modified MIT, internal use exempt, a separate commercial licence only for model-as-a-service operators. Mistral says three more weeks of RL lifted Terminal-Bench 4.0 to 41% from 28.3% on its own mini-SWE-agent run with 30% fewer output tokens, FrontierCode 2.0 to 19.7% (the leading open figure, vendor-run) and the Cyber Index to 51%. Independently, Vals AI measures the open release at 26.26% on Terminal-Bench 4.0 through its standardized mini-SWE-agent harness (27.78% for what this issue reads as the preview checkpoint; Vals figures outside the graded packet), and Artificial Analysis scores the model 26 on Index v5 at \$2.06 per task with Terminal-Bench 4.0 at 25%. One artifact, one benchmark, a 15-point gap between the vendor's run and the measurer's, and the measurer's two readings move 1.5 points the other way across the RL pass the vendor credits with 13: that is the measurer delta, a property of who ran the eval rather than of which ruler the index uses. The API list fell to \$0.70 / \$0.07 cached / \$2.10; Unsloth's 1-bit UD-IQ1_S at 233 GB (a GGUF, the file format llama.cpp loads) keeps 90.6% top-1 agreement; the Hacker News thread reached about 1,700 points. The row moves from closed to open_weights and from medium to high confidence.

SOURCE: MISTRAL AI; HUGGING FACE MODEL CARD AND BLOG

2026-10-27

RELEASE 02

Ant Group (inclusionAI)

OPEN FRONTIER

MOE

Ling-3.1-flash-1019

Ant Group publishes the 560B / 25B-active hybrid-linear MoE under MIT a week after it appeared API-only at 44 on the independent index, resolving last week's deferral

The weights and a Ling-3.1-flash-base checkpoint landed on Hugging Face and ModelScope on October 27 under the MIT licence. The card describes 63 KDA linear-attention layers plus 9 Gated MLA layers (a 7:1 ratio), 512 routed experts with 8 active and 1 shared, a 1M-token context, an FP8 checkpoint that Hugging Face counts at ~572B parameters, and a reference deployment of 8x H200 or 4x GB300 under SGLang. Artificial Analysis scores it 44 on Index v4.3.2 (up from 41 for Ling-3.1-flash) and 29 on the new v5 at \$3.19 per index task, three points behind the best downloadable models (GLM-5.3 and DeepSeek V4.1 Flash 1017, both at 32) and one behind MiMo-V2.6-Pro (30), at an API list of \$0.30 / \$0.90 (leaderboard figures outside the graded packet). Vendor-reported gains (AA-Briefcase ~1,400 to 1,524 Elo, AutomationBench-AA 62% to 68%, CritPt 18% to 26%) are unverified. The first MIT-licensed model at this scale from a Chinese financial group rather than a model lab; placed at high confidence because the artifact, licence and an independent score all exist.

SOURCE: HUGGING FACE MODEL CARD (INCLUSIONAI)

2026-10-27

RELEASE 03

Liquid AI

EDGE / SMALL

DENSE

LFM3 (LFM3-1.7B family)

Six 2B-class models (text, vision, audio, omni, decision) on one 1.7B tri-hybrid backbone trained on 22T tokens, about 1 GB at 4-bit, 256K context, llama.cpp day one

LFM3-1.7B-Base is a 24-layer tri-hybrid of 13 gated short-convolution layers (kernel 7), 5 sliding-window attention layers (256-token window) and 6 global-attention layers, pre-trained on 22T tokens with a final 3.5T-token mid-training stage carrying agentic traces and extending context to 256K. Liquid reports it beating its own LFM2.5-2.6B with 35% fewer parameters and AutomationBench 41.29 against 15.10 for Qwen3.5-9B (GPQA Diamond 63.37 against 80.40), all vendor-run. The text GGUF is 1.02 GiB at Q4_K_M and the Omni-2B GGUF 1.1 GB, with a 1.6 GB KV cache at 128K by Liquid's figure; llama.cpp v0.8.0 (October 28) added the family and new /v1/realtime and /tts endpoints. Weights ship under the LFM Open License v1.0, an Apache 2.0 derivative that revokes commercial rights above \$10M of annual revenue, which the tree records as a commercial-use restriction. The edge-tier question is unchanged: an independent on-device tokens-per-second figure.

SOURCE: LIQUID AI; HUGGING FACE BLOG; LFM OPEN LICENSE

2026-10-28

RELEASE 04

NVIDIA

SPECIALIST

MULTIMODAL

Nemotron 3.5 Transcribe

A 34B / ~4.3B-active promptable speech LLM under OpenMDW-1.1 with a vendor-reported 4.71% mean WER on the Open ASR Leaderboard, one-hour recordings, 40 languages

Transcribe couples a Transformer speech encoder and a frozen Nemotron 3 Diarization model with the Nemotron 3.5 Lightning 30B-A3B hybrid Mamba-Transformer MoE, handling verbatim, written, contextual and speaker-attributed transcription, translation for 31 languages to and from English, language identification and audio question-answering over recordings up to one hour with up to eight speakers in a 65,536-token context; offline only, not streaming. NVIDIA reports the best open-weights mean word error rate on the English Open ASR Leaderboard (4.71%) and 4.62% on the multilingual board, a six-fold cut in named-entity errors with entity context, and up to 2,238 seconds of audio per second on one RTX PRO 6000 Blackwell with NVFP4 and DFlash2, all vendor-run. BF16 runs on one H100 80GB; FP8 (32.3 GiB), NVFP4 (20.1 GiB) and GGUF repositories followed on October 29, alongside the 463,062-hour Granary v2 dataset. The second open audio model of the month after LFM3's omni variant, and the first speech-first row on the Nemotron 3.5 line.

SOURCE: NVIDIA RESEARCH; HUGGING FACE MODEL CARD

2026-10-26

RELEASE 05

Nanbeige LLM Lab (BOSS Zhipin)

SPECIALIST

DENSE

Nanbeige4.3-33B

A dense 33.1B Llama-architecture agentic coding model with a native 524,288-token context surfaces on r/LocalLLaMA tagged Apache 2.0, with no LICENSE file in the repository

The model team at Chinese recruiting platform BOSS Zhipin uploaded Nanbeige4.3-33B and a Base checkpoint on October 24-25 (one to two days before the window); it gained traction from October 26. A dense LlamaForCausalLM with 52 layers, hidden size 7,168, 56 query and 8 KV heads and a 166,912-token vocabulary, native 524,288-token context via rope_theta 1.2e8 with no scaling, thinking and non-thinking modes, XML or JSON tool calling, BF16 weights of ~61.7 GiB and a community Q4_K_M GGUF of ~19.4 GiB; positioned as a teacher for distilling smaller agentic models, with a 3B variant promised. The card tags Apache 2.0 but the repository has no LICENSE file, no training-data or token-count details are published, and the state-of-the-art agentic claim among ~30B open models (JobBench across 35 occupations) is vendor-run. Deferred at grade 3 until the licence file exists.

SOURCE: HUGGING FACE MODEL CARD (NANBEIGE); NVIDIA NEMO AUTOMODEL DOCS

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 **Linear and convolutional layers carry the stack; full attention is one layer in eight**

Ling-3.1-flash-1019 (63 KDA + 9 Gated MLA, 7:1)

Kimi K3.1 Pro (Kimi Delta Attention + Gated MLA, W43)

Nemotron 3.5 Transcribe and Lightning (Mamba-2 hybrid MoE)

LFM3 (13 gated short convolutions, 5 sliding-window, 6 global attention of 24 layers)

Three vendors on three continents shipped the same ratio this week without coordinating: Ant Group's Ling-3.1-flash-1019 runs 63 KDA linear-attention layers (Kimi Delta Attention, a recurrent form whose state does not grow with sequence length) against 9 Gated MLA full-attention layers (multi-head latent attention, which compresses the cached keys and values into a small latent vector), LFM3 runs 13 convolution and 5 sliding-window layers against 6 global ones, and NVIDIA's Transcribe inherits Lightning's Mamba-2-heavy hybrid; Kimi K3.1 Pro's KDA plus Gated MLA backbone is the W43 instance. The consequence for infrastructure is a KV cache that grows with the handful of full-attention layers rather than the depth of the network, which is how a 560B model reaches a 1M-token context on 8x H200 and a 1.7B model carries 128K in 1.6 GB (Liquid's figure). The placement rule holds the hybrids on their primary branch (mixture of experts, decoder-only, multimodal) and records the hybrid attention as an attribute; the case for a branch of their own is stronger on vendor count and infrastructure behaviour than it was in September, and the tree's next quarterly review will revisit it.

SOURCE: HUGGING FACE MODEL CARDS (INCLUSIONAI, LIQUID AI, NVIDIA)

PATTERN 02

One backbone, many heads: a single pretrained core fans out into modality variants and a frozen specialist steers it

LFM3 (six 2B-class text, vision, audio, omni, decision variants on one 1.7B backbone)

Nemotron 3.5 Transcribe (Lightning backbone + frozen Nemotron 3 Diarization)

Genie 4 (Gemini Omni for grounding, a Gemini model as director)

Liquid trained one 1.7B backbone on 22T tokens and shipped six models from it; NVIDIA bolted a speech encoder and a frozen diarization model onto an existing 30B-A3B language MoE and shipped a transcription system; Google's Genie 4 uses Gemini Omni to anchor appearance and a separate Gemini model to direct coherence. The pattern separates the expensive pretraining from the cheap specialisation and makes the modality variant a release-engineering decision rather than a training run, which is why the tree records LFM3 as one row and Transcribe as a child of Super VL rather than as a new family. For buyers it means a vendor's small-model roadmap is now legible from its backbone: the next LFM3 or Nemotron variant will be a head, not a model, and should be evaluated on the head's task rather than re-benchmarked as a new language model.

SOURCE: LIQUID AI; NVIDIA RESEARCH; GOOGLE DEEPMIND

PATTERN 03

Cyber capability as release architecture: the gate moves from training data to distribution

Gemini 4 Carbon (Fairwind defender-only rollout, no GA date)

GPT-6.1 Terra (Critical cybersecurity classification, GPT-6 Astra safeguards)

Mistral Large 4 (state-authority red-teaming; offensive-security variant restricted to vetted partners)

Kimi K3.1 Pro (cyber data removed from mid-training and RL, W43)

Four labs in two weeks have treated offensive-cyber capability as the variable that decides how a general-purpose model ships. Moonshot removed data; Mistral red-teamed with state authorities, observed 'a potential for uplift to conduct real attacks' in live traffic, and kept an offensive variant for vetted partners; OpenAI ships Terra at its Critical tier with a safeguards stack; Google ships the independent leader to defenders only. The open question this publication cannot yet answer is whether offensive-cyber capability is separable from general long-horizon agentic capability (Interconnects raised it; Kimi's independent Cyber Index score of 64% against its curb claim is the one datapoint, and it cuts against separability). If it is separable, Carbon's gate is a marketing cost; if not, every frontier release becomes a defender-first allocation and open weights are structurally capped below the gate. The AI Stack Weekly holds the question open and bets 38% on Carbon's general availability by December 15.

SOURCE: GOOGLE; OPENAI SYSTEM CARD ADDENDUM; MISTRAL AI

PATTERN 04

The ruler is the architecture: private test sets, mergeability grading and standardized harnesses re-base the leaderboard

Artificial Analysis Intelligence Index v5 (9 evaluations, 60% private)

Cognition FrontierCode 2.0 (132 maintainer-built tasks graded on mergeability)

Vals AI Terminal-Bench 4.0 (standardized mini-SWE-agent harness with cost per test)

Three independent measurers changed what they measure in one week. Artificial Analysis moved 60% of its index to private test sets and added FrontierCode 2.0 and Terminal-Bench-Science, which re-based every score and widened the open-weights gap from 3 to 12 points; Cognition's FrontierCode 2.0 grades 132 maintainer-built tasks on whether a maintainer would merge the result rather than whether tests pass; Vals AI runs every model through one mini-SWE-agent harness and publishes cost per test alongside the score. Each is an improvement and each is a methodology choice: Vals's standardized harness scores Terra at 49.49% where OpenAI's own harness scores 50.3% and the tracker's 55.6%, Artificial Analysis's own Stirrup agent scores FrontierCode 2.0 differently from Cognition's run (DeepSeek V4.1 Flash at 28% and Terra at 42% on Stirrup against Cognition's table; Stirrup figures outside the graded packet), and Mistral Large 4 moves 15 points on Terminal-Bench 4.0 between the vendor's run and Vals's. The benchmark table in this issue carries the ruler next to every number; an architect comparing two scores should first check that they came from one instrument.

SOURCE: ARTIFICIAL ANALYSIS; COGNITION; VALS AI

BENCHMARK MOVES

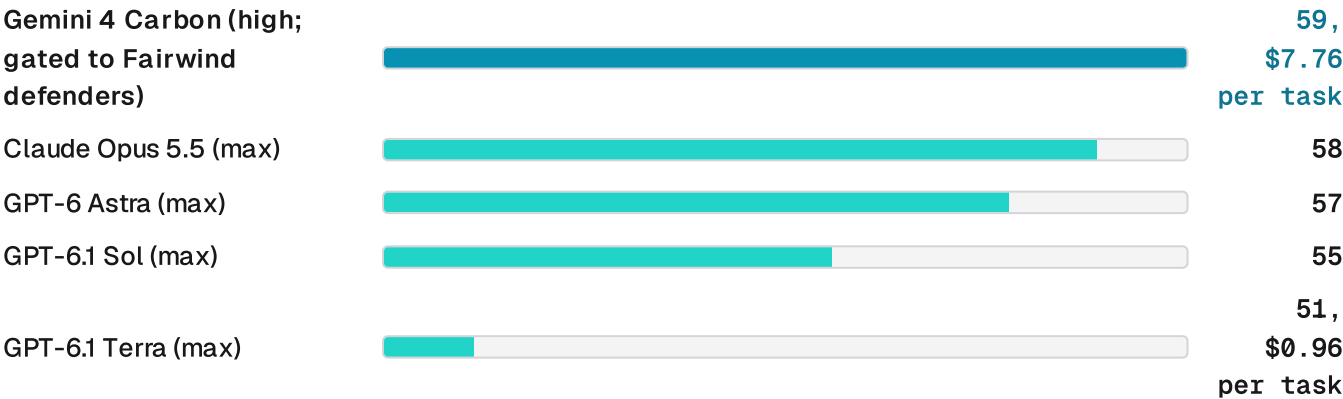
Where the leaderboard moved.

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

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V5 (LAUNCHED OCTOBER 27; 9 EVALUATIONS, 60% PRIVATE TEST SETS)

MOVE 01

A new ruler, not a new result: v5 re-based every score and every cost-per-task figure, Gemini 4 Carbon entered first at 59 on October 28, and the open-weights gap to the frontier widened from 3 points (v4.3.2) to 12. Scores below are v5 at each model's highest listed effort; none is comparable to last week's table.

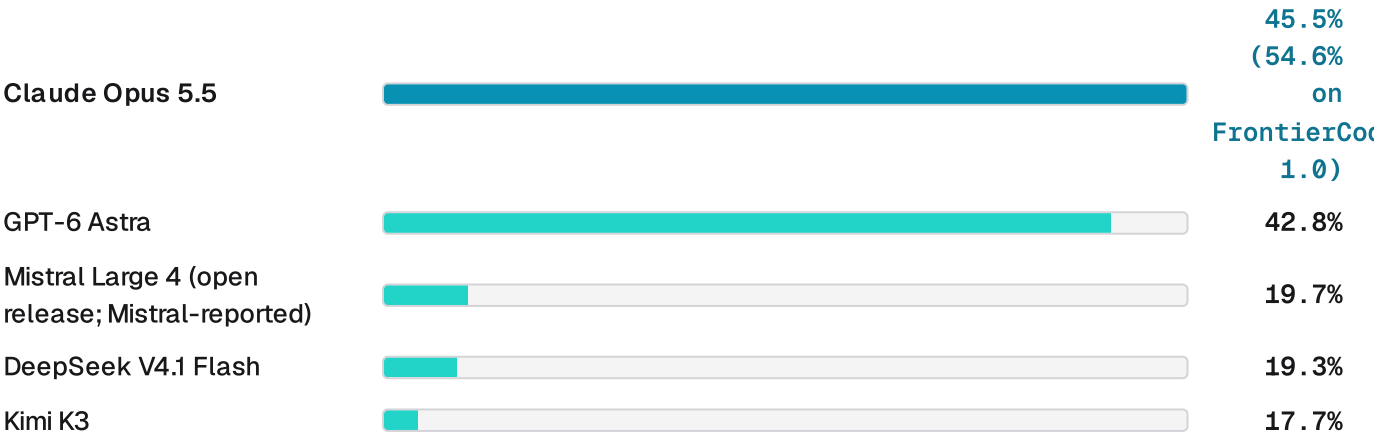


SOURCE: ARTIFICIAL ANALYSIS

COGNITION FRONTIERCODE 2.0 (132 MAINTAINER-BUILT TASKS GRADED ON MERGEABILITY)

MOVE 02

The first mergeability-graded coding benchmark puts the frontier below half and, on Cognition's harness, every open model below a fifth; Claude Opus 5.5 leads at 45.5%, down from 54.6% on FrontierCode 1.0, and the best open figure is Mistral's own 19.7% for Large 4 (vendor-run). Artificial Analysis's Stirrup agent runs the same benchmark and scores DeepSeek V4.1 Flash at 28% and GPT-6.1 Terra at 42%, so the open-closed gap here moves with the instrument.

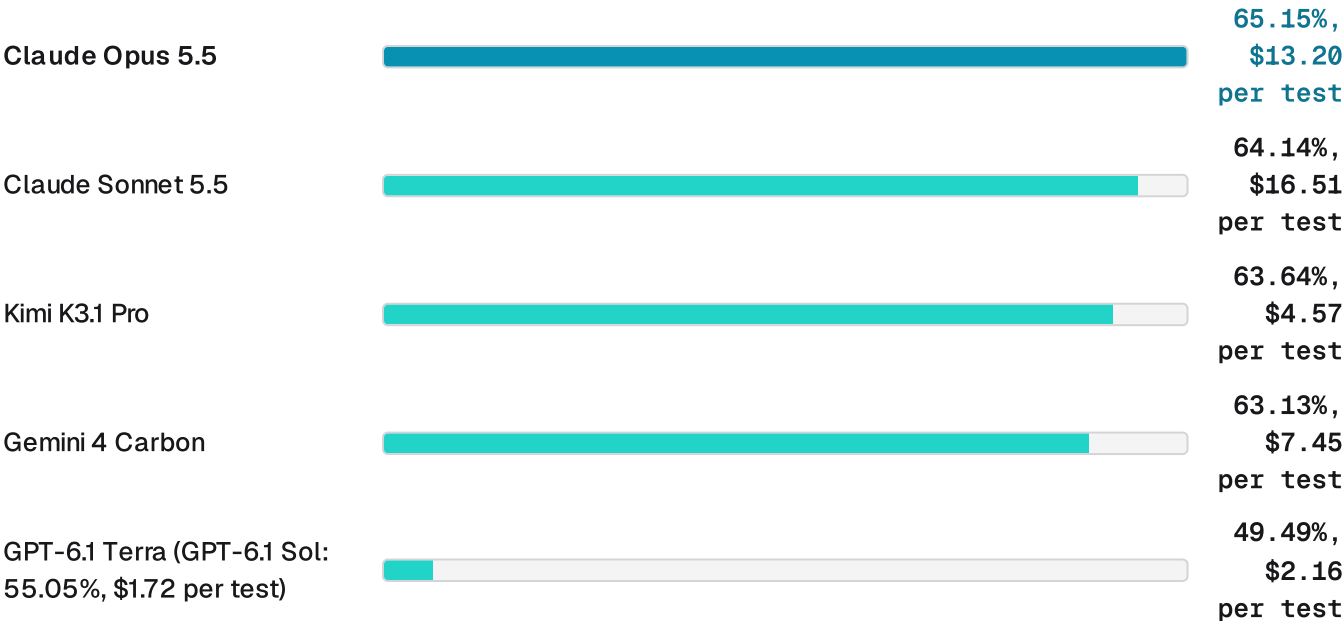


SOURCE: COGNITION; MODELCOMPARE

VALS AI TERMINAL-BENCH 4.0 (STANDARDIZED MINI-SWE-AGENT HARNESS, COST PER TEST)

MOVE
03

On one harness with one cost basis the top of the table compresses: Opus 5.5, Sonnet 5.5, Kimi K3.1 Pro and Carbon sit within two points of each other at 63-65%, with Sonnet costing more per test than Opus on this run, and the \$1 / \$5 Terra reaches 49.49% at \$2.16 per test against \$13.20 for Opus 5.5. The same harness scores Terra four points below OpenAI's own figure and six below the Artificial Analysis harness. Rows are limited to five; the Mistral Large 4 figures from this harness sit in the exhibit below. Vals figures were read from the leaderboard after the research stage and are outside the graded packet except where the packet carries them (Terra, Sol, Kimi, Carbon).

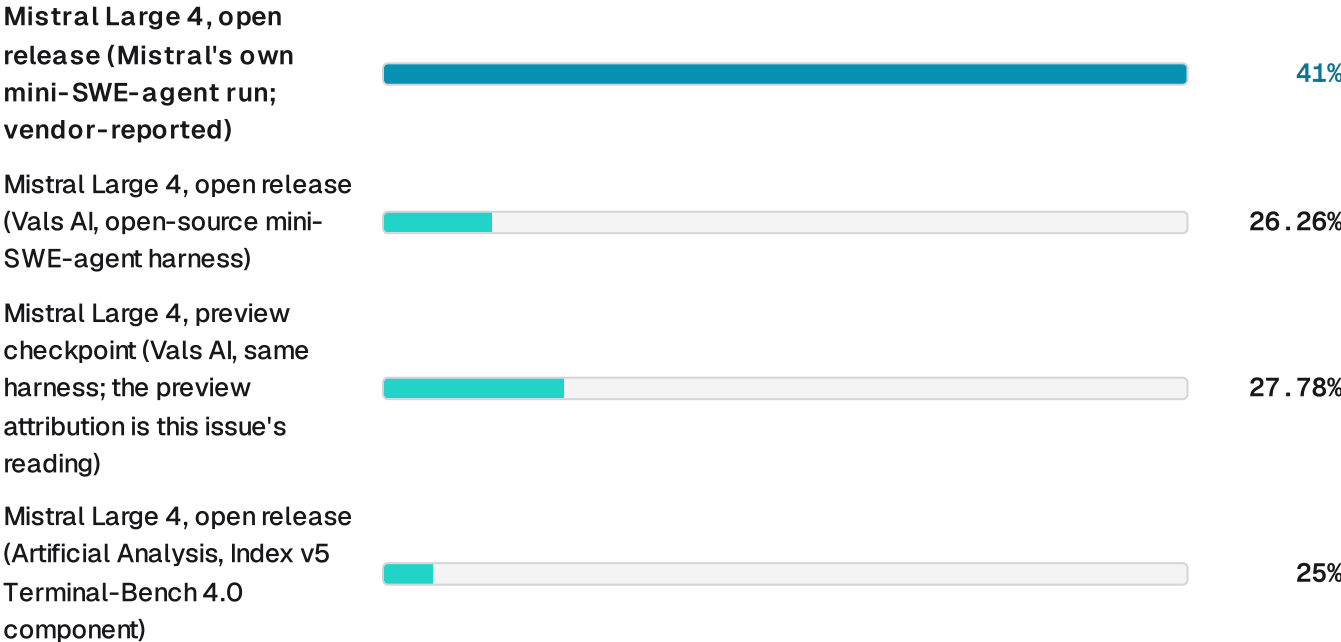


SOURCE: VALS AI

MISTRAL LARGE 4 ON TERMINAL-BENCH 4.0, BY MEASURER (THE WEEK'S MEASURER-DELTA EXHIBIT)

MOVE
04

One artifact, one benchmark, one agent scaffold, two measurers: Mistral reports 41% for the open release from its own mini-SWE-agent run (up from 28.3% on the preview's vendor figure, with 30% fewer output tokens), while Vals AI's standardized mini-SWE-agent harness measures the same open weights at 26.26% and what this issue reads as the preview checkpoint at 27.78%; Artificial Analysis's v5 run has the open release at 25%. The 15-point spread between the vendor and the independent measurers on the same weights is the measurer delta, and the direction of travel is the sharper point: the vendor credits its RL pass with 13 points, the independent harness moves 1.5 points the other way across the same two checkpoints. (Vals figures outside the graded packet.)



SOURCE: MISTRAL AI; VALS AI; ARTIFICIAL ANALYSIS

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class.
As of Oct 31, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	Claude Opus 5.5 (procurable leader)	Gemini 4 Carbon (independent leader at 59; gated to Fairwind defenders)	The row ranks what a buyer can procure. Carbon leads Artificial Analysis's v5 index at 59 against Opus 5.5's 58 and leads LMArena Text at 1,528, and is served only to vetted cyber defenders with no general-availability date, so it enters as the challenger with the asterisk the row gave Argon during its thirteen gated days in September. Opus 5.5 holds on v5 (58), FrontierCode 2.0 (45.5%, first) and Vals Terminal-Bench 4.0 (65.15%, first). The row flips to Carbon the week it reaches the paid Gemini API tier (p147, 38% by December 15); Kimi K3.1 Pro (W43 challenger) drops to third pending its weights date.
Open frontier	GLM-5.3 and DeepSeek V4.1 Flash 1017 (tied at 32)	MiMo-V2.6-Pro	Re-based to v5 and the order flips, on leaderboard figures read after the research stage (outside the graded packet): GLM-5.3 (max) and DeepSeek V4.1 Flash 1017 (max) are tied at 32 as the best models a buyer can download, at \$5.01 and \$0.60 per index task respectively, with MiMo-V2.6-Pro at 30 (GLM and MiMo swap places from v4.3.2, where MiMo led 46 to 45), Ling-3.1-flash-1019 and MiniMax-M3.1-Flash at 29, and Mistral Large 4 entering at 26. The row carries both leaders because the tie is on the index and the cost gap is eight-fold in DeepSeek's favour. The row ranks downloadable weights, so Kimi K3.1 Pro's 46 (API-only, weights unreleased; Moonshot's stated date is November 4) is carried as the tier's gated anchor rather than its leader, the same treatment the closed-frontier row gives Carbon. Two artifacts entered the tier this week under MIT-family licences (Ling-3.1-flash-1019; Mistral Large 4 under a modified MIT); the Weekly's p150 bets at 46% that a downloadable model other than Kimi reaches 40 on v5 by year-end.

TIER	LEADER	CHALLENGER	READ
Reasoning	Claude Opus 5.5	GPT-6 Astra	Re-based to v5: Opus 5.5 (max) at 58 and Astra (max) at 57 are the two highest procurable scores on the new index, with GPT-6.1 Sol at 55 and Sonnet 5.5 (W43's challenger) not yet listed at its max setting on v5 in the packet. Carbon at 59 would lead and is gated (the closed-frontier row carries it). Terminal-Bench-Science, new to v5, is the component to watch for the reasoning tier specifically.
Coding	Claude Opus 5.5	Kimi K3.1 Pro	Flipped from Sonnet 5.5 on the row's standing rule (a change needs a second benchmark or a second week): Opus 5.5 leads Cognition's FrontierCode 2.0 at 45.5%, Vals AI's standardized Terminal-Bench 4.0 at 65.15%, and Artificial Analysis's Coding Agent Index with Claude Code at 61.7, three independent coding measurements in one week. K3.1 Pro holds the challenger slot on a second harness (Vals 63.64%, 1.5 points behind Opus and half a point behind Sonnet 5.5's 64.14%) after last week's 67% on the tracker's; the top four on Vals sit inside two points, so the row's order is a judgment on breadth of measurement, not a gap in capability. Carbon's 63.13% on Vals and 59.4 on the Coding Agent Index would make it the challenger if it were procurable. Mistral Large 4's 19.7% on FrontierCode 2.0 is the open high-water mark on Cognition's harness and is vendor-run.
Multimodal	Gemini 4 Argon	Gemini 3.8 Live Extended Thinking	Held. Carbon (text, image, audio, video in) would take the row on general availability and is gated; Genie 4 is a world model excluded from the tree, and its use of Gemini Omni for grounding is the architecture-watch item. The open side added two audio-capable rows (LFM3's omni variant, Nemotron 3.5 Transcribe) with vendor-only numbers.

TIER	LEADER	CHALLENGER	READ
Sub-dollar hosted	Claude Haiku 5.5	Ling-3.1-flash-1019	Challenger changes, on v5 figures: Ling-3.1-flash-1019 at \$0.30 / \$0.90 per million scores 29 on v5 at \$3.19 per index task, level with MiniMax-M3.1-Flash (29 at \$0.15 / \$0.60), replacing GPT-6 Luna (W43's challenger, not yet carried at its max setting on v5 in the packet), with MIT weights behind the API; the tie with MiniMax is broken on the downloadable weights. GPT-6.1 Terra at \$1.00 input sits on the row's boundary and is not sub-dollar; Mistral Large 4's new \$0.70 / \$2.10 list qualifies on input and scores 26 on v5 at \$2.06 per task, cheaper per task than Ling, so it is carried as the open alternative. DeepSeek V4.1 Flash 1017, tied for the open-frontier lead at 32 and \$0.60 per index task, would take this row's challenger slot on per-task cost if its list price qualifies; the packet does not carry its list price, so it is noted here and not ranked until it does. Per-task cost is now the row's second sort key.
Edge / small	LFM2.5-24B-A2B	Nemotron 3.5 Bolt 74B-A9B	Held on the standing rule. LFM3 (1.7B backbone, about 1 GB at 4-bit, six variants) is the week's edge release and every number on it is Liquid's; Flint (W43) and Bolt E4B (W43) are likewise waiting for an independent run. The row's leader changes when an independent harness scores any of the three; Nemotron 3.5 Transcribe is workstation-class (one H100 in BF16) and is recorded in the specialist tier rather than here.

VENDOR SIGNALS

Pricing, gating, depreciation.

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

2026-10-26

SIGNAL 01

OpenAI

GPT-6.1 Terra lists \$1 / \$0.05 cached / \$5 per million with a 2x input and 1.5x output multiplier above 272K tokens; available in Codex and ChatGPT Work, not consumer ChatGPT

Two signals in one rate card: the cheapest cached input the tree carries for a frontier-adjacent model (\$0.05, half of GPT-6.1 Sol's), and a tiered long-context surcharge that makes the 1.05M context a different product above 272K. The Codex-and-Work-only distribution (The Application Layer carries the custom-GPT change) says Terra is a computer-use tier sold to agents, not a chat model.

SOURCE: OPENAI API DOCS

2026-10-28

SIGNAL 02

Google

Gemini 4 Carbon served only through the Fairwind Program (650+ vetted defenders) with no general-availability date; \$2 / \$10 introductory pricing with a reversion to \$4 / \$20

A distribution gate on the independent leader, priced as if it were generally available. Argon's identical Fairwind-first launch opened to the paid tier in thirteen days; if Carbon's does not, the introductory rate's reversion clock starts before most buyers can call the model, and the cyber gate becomes the release architecture rather than a launch phase.

SOURCE: GOOGLE; GOOGLE DEEPMIND FAIRWIND PROGRAM

2026-10-29

SIGNAL 03

Mistral AI

Mistral Ecosystem Open-Weight License (modified MIT; separate commercial licence only for model-as-a-service operators), API cut to \$0.70 / \$2.10, offensive-security variant restricted to vetted partners

The narrowest gate of the autumn's open releases: internal use is exempt and the only carve-out is resale as a service, against MiniMax's and Muse's \$20M revenue gates and Liquid's \$10M. The 49% API cut on the same day the weights shipped is the price of competing with your own download, and the vetted-partner offensive variant is the cyber gate applied to an open model.

SOURCE: MISTRAL AI; HUGGING FACE LICENSE

2026-10-27

SIGNAL 04

Liquid AI

LFM3 ships under LFM Open License v1.0, an Apache 2.0 derivative that revokes commercial rights above \$10M of annual revenue

The same gate the LFM2.5 line carried, now on a family built for device deployment, which is where the gate bites hardest: a handset or appliance vendor above \$10M of revenue needs a separate licence to ship a 1 GB model. The tree records the restriction on the row; buyers should read the licence before the benchmark table.

SOURCE: LIQUID AI LFM OPEN LICENSE

2026-10-26

SIGNAL 05

Nanbeige LLM Lab (BOSS Zhipin)

Nanbeige4.3-33B card tags Apache 2.0 with no LICENSE file in the repository; no training-data or token-count disclosure

A licence tag without a licence text is not a licence, and the tree defers the row until the file exists. The model is otherwise the kind of artifact the edge and coding tiers want (dense 33B, native 512K context, 19.4 GiB at Q4_K_M), which is why the deferral is recorded rather than the model ignored.

SOURCE: HUGGING FACE MODEL CARD (NANBEIGE)

WATCHLIST

On the radar next.

5 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.

NOVEMBER 4

WATCH 01

Kimi K3.1 Pro weights date

Moonshot's stated date for full weights; the row moves from closed to open and the open-frontier tier re-ranks if the artifact reproduces the API score (46 on v5; 55 on v4.3.2), and the ledger's p142 (58%) tests whether the licence carries a K3-style gate. A slip restores the label-versus-artifact pattern this week interrupted.

BY DECEMBER 31

WATCH 02

A downloadable model other than Kimi K3.1 Pro at 40 or above on Artificial Analysis v5 (a DeepSeek or Qwen refresh, or a GLM-5.3 successor)

The best downloadable scores on v5 are GLM-5.3's and DeepSeek V4.1 Flash 1017's 32 against Kimi K3.1 Pro's API-only 46; the Weekly's p150 bets at 46% that a downloadable model other than Kimi reaches 40 by year-end (Kimi is excluded so the bet does not resolve on the November 4 weights date), which decides whether the 12-point gap v5 opened is holding for everyone but Moonshot. The open-frontier row re-ranks on each new v5 open score.

BY DECEMBER 15

WATCH 03

Gemini 4 Carbon on the paid Gemini API tier

The closed-frontier, reasoning and multimodal rows all flip on general availability, and The AI Stack Weekly's cache-read lever flips on Carbon's \$0.10 cached input; p147 holds it at 38%, so the base case is that the gate holds through the window.

DECEMBER 1

WATCH 04

Gemini 4 Argon introductory-rate reversion to \$4 / \$20

Third-party write-ups date the reversion to December 1 and Google has not; if it reverts, the \$2 / \$10 band loses a member and the Weekly's mid-tier lever count falls to three, and Carbon's identical introductory structure inherits the same clock.

NOVEMBER

WATCH 05

Genie 4 technical report and Nanbeige LICENSE file

Genie 4 stays excluded until Google publishes resolution, frame rate or a report (the director pattern would then be assessable); Nanbeige4.3-33B's row is created the day a LICENSE file appears in the repository.

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, 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 28

Week 44 of 2026 · October 31, 2026

WEB

brianletort.ai/industry/model

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