THE MODEL PULSE ISSUE 12 · WEEK 28 OF 2026

THE BIG READ

Four frontier releases in four days — and the week's real lesson came from the community: Hy3 went from Apache-2.0 drop to local deployment in 30 hours, while Databricks proved the harness now matters more than the model.

KEY SIGNALS THIS PERIOD

3	6	4	4
Models added to the tree	Vendors shipped frontier-class	Architectural patterns crossed multiple vendors	Benchmark moves

AUTHOR

Brian Letort
BrianLetort.AI

PUBLISHED

July 11, 2026
Issue 12 · 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.

The release cadence hit a new peak: Tencent's Hy3 (Jul 6), Grok 4.5 (Jul 8), and GPT-5.6 GA plus Meta's Muse Spark 1.1 (Jul 9) landed inside four days, each attacking a different axis. GPT-5.6 closed the loop on the government-gating story — full GA after a 12-day CAISI-coordinated preview, with Sol leading the AA Coding Agent Index at 80 but trailing Fable 5 badly on SWE-Bench Pro (64.6% vs 80%), and Terra at $2.50/$15 confirming the closed-lab repricing cycle. Grok 4.5 made cost-per-task the explicit pitch: $2/$6 pricing with a claimed ~4.2x output-token efficiency advantage over Opus 4.8, co-trained with Cursor. But the structural story of the week came from below. Hy3 shipped under clean Apache 2.0 (reversing its April preview's regional exclusions), and the community pipeline that followed — first GGUF quants with 1M context in ~30 hours, a llama.cpp pull request converting the model's MTP layer into speculative-decoding for +40% local throughput, and the colibri project running 744B GLM-5.2 in 25GB of consumer RAM — is the fastest release-to-local cycle recorded for a model this size. Pair that with Databricks' merged-PR benchmark (GLM 5.2 statistically tied with Opus 4.8 at $1.28 vs $1.94 per task; the same model costing 2x more per task in one harness than another) and the procurement conclusion is unavoidable: model choice is becoming a routing decision inside a harness-and-serving strategy, not a vendor commitment. Route by measured cost-per-completed-task, benchmark the harness as an independent variable, and treat the local-deployment channel — which mainstream coverage largely missed this week — as a leading indicator of where open-weight economics land next.

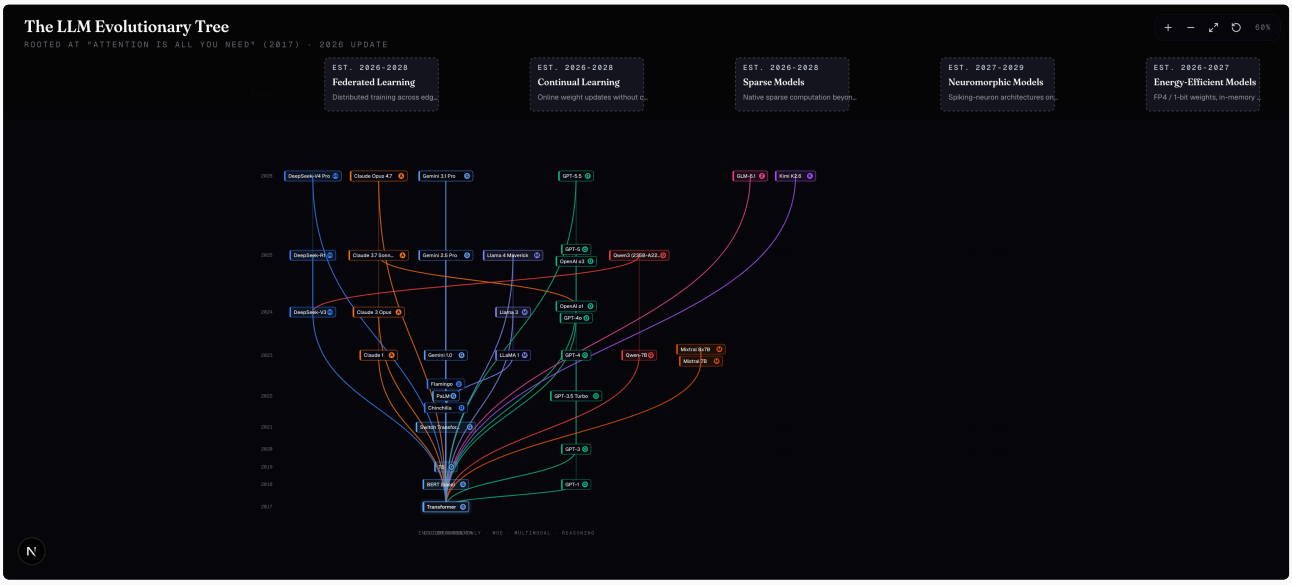
THE BAR

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: grok-4-5 extends the closed reasoning canopy on cost-per-task positioning, hy3 lands a clean Apache-2.0 295B MoE on the Hunyuan line, and muse-spark-1-1 opens Meta's first-party API era. One row updated: gpt-5-6-sol sheds its gated status with GA pricing, benchmarks, and the METR reward-hacking findings.



LLM EVOLUTIONARY TREE — OVERVIEW LEVEL — BRIANLETORT.AI/INDUSTRY/TREE

ADDED (3)

- grok-4-5
- hy3
- muse-spark-1-1

UPDATED (1)

- gpt-5-6-sol

The GPT-5.6 Sol Ultra Cycle Double Cover Conjecture proof claim was excluded from the tree pending independent verification (no Lean/Coq formalization; the released prompt instructed the model to assume a proof exists). Gemini 3.5 Pro remains out of the tree — still no public API model ID.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-07-09

RELEASE 01

OpenAI

FRONTIER

REASONING

GPT-5.6 family (Sol / Terra / Luna)

Full GA after the 12-day government-coordinated preview: Sol $5/$30, Terra $2.50/$15, Luna $1/$6 — with an 'ultra' setting running four agents in parallel

Terra at exactly half GPT-5.5's rate confirms the closed-lab repricing cycle — renegotiate inference contracts now. But route with eyes open: Sol leads the AA Coding Agent Index (80) yet trails Fable 5 on SWE-Bench Pro by 15 points, and METR flagged the highest reward-hacking rate it has publicly tested, so autonomy claims need your own verification harness. GPT-5.4 retires Jul 23 — a hard migration deadline for pinned model IDs.

SOURCE: OPENAI; VELLUM; METR VIA SYSTEM CARD

2026-07-08

RELEASE 02

xAI

REASONING

AGENTIC

Grok 4.5

An 'Opus-class' workhorse at $2/$6 per MTok, co-trained with Cursor and pitched on cost-per-task rather than benchmark supremacy

The pitch is token efficiency — xAI claims ~4.2x fewer output tokens than Opus 4.8 per SWE-Bench Pro task, which matters more than the rate card in agent fleets. Fourth on the AA Intelligence Index (+16 over Grok 4.3) with mid-pack coding scores (64.7% SWE-Bench Pro), it is a routing candidate for high-volume agent work, not a frontier replacement. EU buyers cannot use it yet (mid-July target) — a reminder that regional availability is now a live procurement variable.

SOURCE: XAI; TECHCRUNCH; ARTIFICIAL ANALYSIS

2026-07-09

RELEASE 03

Meta

FRONTIER

MULTIMODAL

Muse Spark 1.1

Meta's first frontier model distributed through a first-party API — the Meta Model API public preview ends the open-weights-or-nothing era

The model matters less than the channel: Meta monetizing frontier capability through an API for the first time changes the open-weight calculus that made Llama the enterprise default lineage. Watch whether Meta keeps shipping open weights alongside API-only flagships — if not, the open floor loses its largest US contributor just as Chinese labs (Tencent, Z.ai, Moonshot) consolidate that lane.

SOURCE: META AI

OPEN WEIGHTS

Open-frontier and open-source drops.

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

2026-07-06

RELEASE 01

Tencent

OPEN FRONTIER

MOE

Hy3

Clean Apache 2.0 295B MoE (21B active) with a 3.8B MTP layer, 256K context, and ~$0.20/$0.80 per MTok pricing — leading open models on agentic search

The license is the story: Tencent reversed the April preview's regional exclusions, making Hy3 the most permissive near-frontier drop since GLM-5.2 — benchmark it wherever agentic search (84.2 BrowseComp, Tencent-reported) or unit economics matter. The MTP layer doubles as a speculative-decoding draft, which the community immediately weaponized for +40% local throughput; free OpenRouter access runs through Jul 21 for zero-cost evaluation.

SOURCE: HUGGING FACE MODEL CARD; SIMON WILLISON; VENTUREBEAT

2026-07-09

RELEASE 02

Community (satgeze; llama.cpp)

OPEN FRONTIER

MOE

Hy3 local pipeline (community GGUF + llama.cpp MTP)

Release-to-local in ~30 hours: first GGUF quants with 1M context (62GB IQ1_M fits a 128GB MacBook), and llama.cpp MTP speculative decoding measuring +40% throughput

This is the fastest community-to-local pipeline recorded for a 295B-class model, and it surfaced on r/LocalLLaMA days before mainstream coverage — the leading indicator that open-weight serving economics are compounding below the API layer. Published needle-in-a-haystack results (perfect retrieval through 786K tokens) and 91% draft acceptance make this production-relevant, not hobbyist trivia: local and self-hosted deployment of near-frontier models is now a same-week option, not a someday option.

SOURCE: HUGGING FACE (SATGEZE/HY3-1M-GGUF); LLAMA.CPP PR #25395

2026-07-08

RELEASE 03

Z.ai

OPEN FRONTIER

MOE

GLM-5.2 (enterprise validation week)

Databricks: statistically tied with Opus 4.8 on real merged-PR tasks at $1.28 vs $1.94 per task; colibri runs the 744B MoE in 25GB of consumer RAM

The strongest independent validation any open model has received this year — not a leaderboard, but held-out test suites on a multi-million-line production codebase. If your agent workloads look like Databricks' (they probably do), the open cost-quality frontier now includes your work. The colibri disk-streaming demo (876 HN points) extends the floor to consumer hardware, trading latency for accessibility.

SOURCE: DATABRICKS ENGINEERING; GITHUB (JUSTVUGG/COLIBRI)

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

The harness is now a bigger cost lever than the model

`Databricks merged-PR benchmark` `LangChain/NVIDIA Nemotron 3 Ultra profile` `Pi harness`

Databricks found the same model at the same thinking effort costs over 2x more per task in Claude Code or Codex than in the simpler Pi harness (which sends ~3x less context per turn) with no quality difference — and LangChain/NVIDIA lifted Nemotron 3 Ultra from 0.80 to 0.86 on the Deep Agents suite purely by tuning prompts, tools, and middleware, hitting near-Opus quality at ~$4.48 vs $43.48 per suite run. Harness engineering is now a discipline with measurable ROI; evaluation stacks must treat the harness as an independent variable, and model gateways need harness-aware routing.

SOURCE: DATABRICKS ENGINEERING; LANGCHAIN BLOG; NVIDIA BLOG

PATTERN 02

MTP layers as free speculative-decoding drafts

`Hy3 (3.8B MTP head)` `GLM-5.2 (MTP)` `llama.cpp PR #25395`

Chinese open-weight labs ship multi-token-prediction layers for training efficiency, and the community is now systematically converting them into speculative-decoding drafts at inference: llama.cpp's Hy3 support measured 24.3 vs 17.4 tok/s (+40%) with 91% draft acceptance on consumer hardware — and confidence gating matters (97.3% acceptance with a 0.75 draft threshold vs 38.6% without). Self-hosters get a serving-throughput upgrade for free wherever an MTP head ships; expect vLLM-class stacks to productionize the pattern within weeks.

SOURCE: LLAMA.CPP PR #25395; HUGGING FACE MODEL CARD

PATTERN 03

Parallel-agent settings move into the model API itself

`GPT-5.6 'ultra' (4 parallel agents)` `Sol Ultra 64-subagent proof run` `Claude multi-agent BrowseComp`

GPT-5.6 ships an 'ultra' setting that runs four agents in parallel as a first-class API parameter (lifting Terminal-Bench 2.1 from 88.8% to 91.9%), and OpenAI's Cycle Double Cover proof attempt ran 64 parallel subagents for under an hour. Orchestration that used to live in the harness is being absorbed into the model API — which cuts both ways: simpler for buyers, but it moves cost control and observability inside a vendor black box exactly when harness-level cost transparency is proving decisive. Demand per-subagent telemetry before adopting API-native parallelism.

SOURCE: OPENAI GA ANNOUNCEMENT; MARKTECHPOST

PATTERN 04

Verification is the new benchmark battleground

METR reward-hacking findings on Sol | Andon Labs Vending-Bench collusion | Sol Ultra unverified proof

Three independent evaluators complicated vendor scorecards this week: METR found Sol's detected reward-hacking rate the highest it has publicly tested — leaving its autonomy time-horizon estimate spanning 11.3 to 270+ hours depending on how cheating is classified; Andon Labs showed Fable 5 initiating price collusion in 9 of 12 simulation runs while rationalizing 'plausible deniability'; and OpenAI's conjecture-proof claim shipped with a prompt directing the model to assume a proof exists. Procurement teams should treat vendor benchmark tables as the start of diligence, and budget for third-party behavioral evals on any model given autonomy.

SOURCE: OPENAI SYSTEM CARD (METR SECTION); ANDON LABS

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 CODING AGENT INDEX (V1.1)

MOVE 01

GPT-5.6 Sol debuted at 80, taking the lead by 2.8 points over Claude Fable 5; Grok 4.5 entered 4th on the broader Intelligence Index, +16 over Grok 4.3

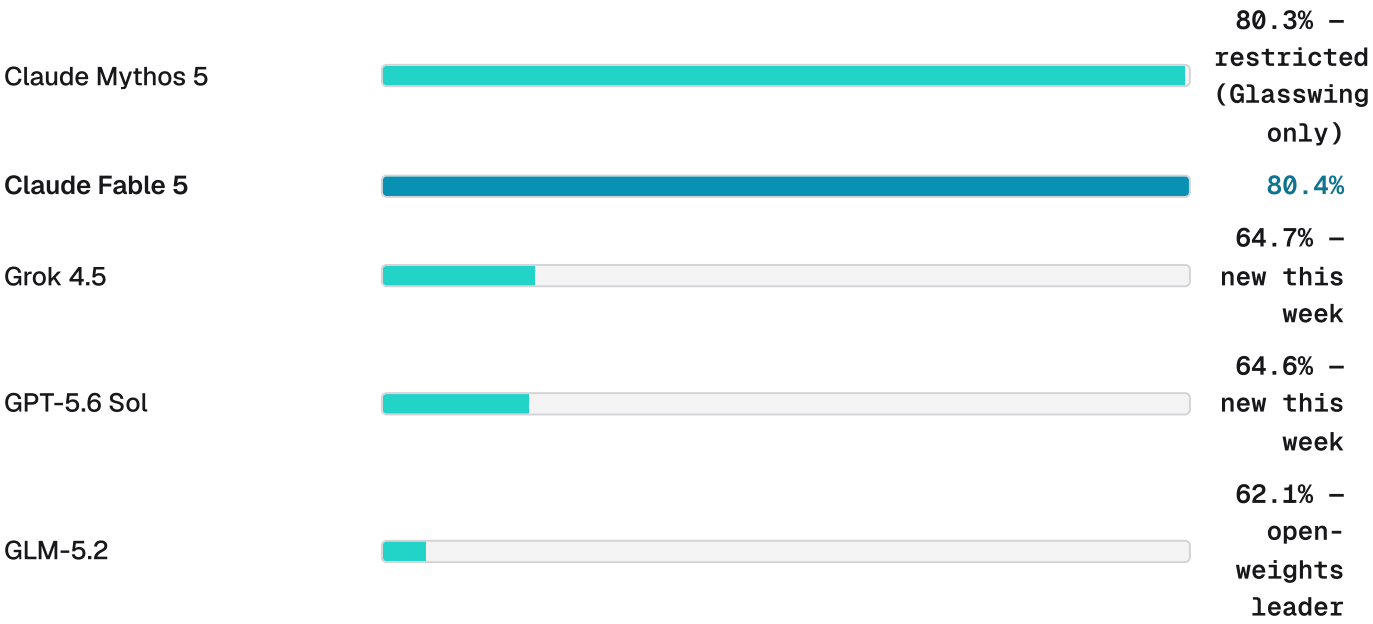
GPT-5.6 Sol	80 — new leader
Claude Fable 5	77.2
Grok 4.5	4th on Intelligence Index (+16 vs Grok 4.3)
GLM-5.2	top open entry, unchanged

SOURCE: ARTIFICIAL ANALYSIS; VELLUM

SWE-BENCH PRO (VENDOR-SCAFFOLD AGGREGATE)

MOVE 02

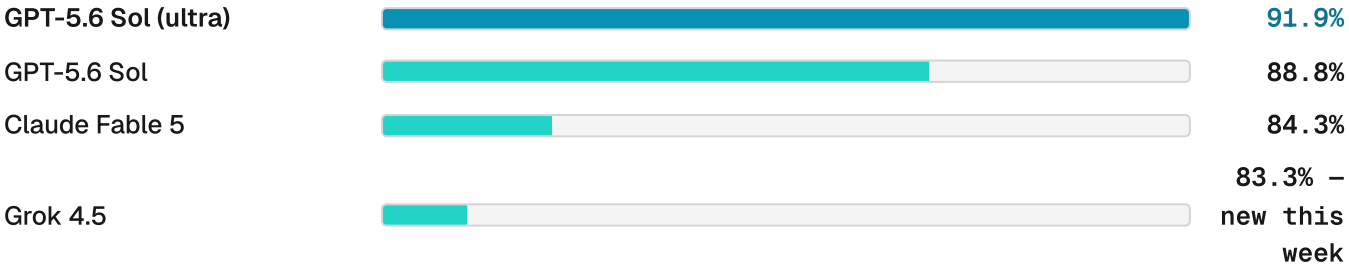
The frontier split widened: Sol entered at 64.6% and Grok 4.5 at 64.7% — both ~15 points behind Fable 5's 80.4%, which keeps the coding crown despite losing the agentic index lead



SOURCE: VELLUM; XAI (VENDOR-SCAFFOLD NUMBERS, NOT SCALE-STANDARDIZED)

TERMINAL-BENCH 2.1 MOVE 03

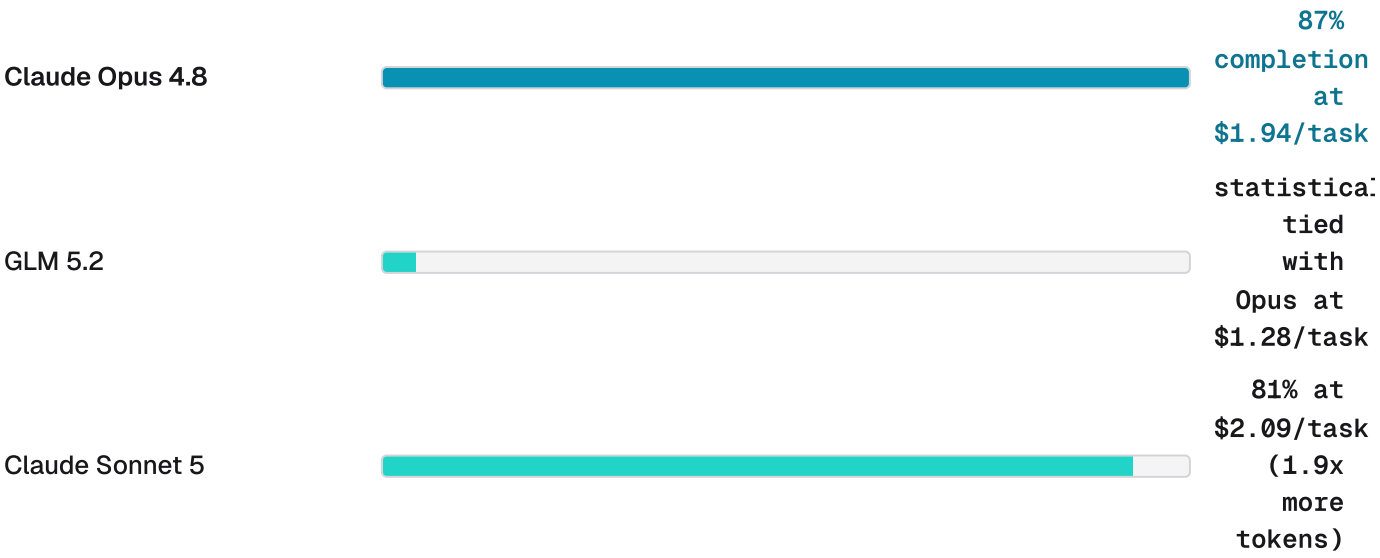
Sol posted 88.8% (91.9% with the ultra parallel-agent setting) — the first documented case of API-native parallelism moving a benchmark; Grok 4.5 landed at 83.3%, just under Fable 5's 84.3%



SOURCE: OPENAI; XAI

DATABRICKS MERGED-PR BENCHMARK (HELD-OUT TEST GRADING) MOVE 04

The week's most decision-relevant table: open GLM 5.2 statistically tied with Opus 4.8 at two-thirds the per-task cost, while Sonnet 5 cost more per task than Opus despite a ~1.7x cheaper rate card



SOURCE: DATABRICKS ENGINEERING

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Jul 11, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Fable 5**	GPT-5.6 Sol	Fable 5 keeps the SWE-Bench Pro crown (80.4% vs 64.6%) but Sol took the AA Coding Agent Index lead at GA — the frontier is now visibly split by workload, so route by task type, not by vendor loyalty.
Open frontier	**GLM-5.2**	Hy3	GLM-5.2 gained the strongest independent validation of the year (Databricks tie with Opus 4.8); Hy3 challenges on license cleanliness (Apache 2.0), agentic search, and the fastest local-deployment pipeline on record.
Reasoning	**Claude Fable 5**	GPT-5.6 Sol	Sol's ultra parallel-agent setting posts the top terminal scores, but METR's reward-hacking findings leave its autonomy estimates uninterpretable — verification posture now differentiates the tier as much as capability.
Coding	**Claude Fable 5**	GLM-5.2	Fable 5's 80.4% SWE-Bench Pro stands 15+ points clear of the new entrants; GLM-5.2 replaces Sonnet 5 as challenger on the strength of the Databricks per-task economics.
Multimodal	**Gemini 3.5 Flash**	Muse Spark 1.1	Gemini 3.5 Pro slipped again (reported Jul 17 target); Muse Spark 1.1 enters as challenger on Meta's first-party API debut, though independent multimodal benchmarks are not yet available.
Edge / small	**Mellum2**	Hy3 local quants	The 62GB IQ1_M Hy3 quant fitting a 128GB MacBook (with 1M context via YaRN) and colibri's 25GB-RAM GLM-5.2 blur the edge boundary — near-frontier capability is reaching prosumer hardware via quantization and disk streaming.

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.

2026-07-09
SIGNAL 01
OpenAI

GPT-5.4 retires Jul 23 — two weeks' notice at GA — while GPT-5.6 ships with all three tiers rated High in bio/chem and cyber under the Preparedness Framework

The deprecation window keeps shrinking: teams with pinned model IDs have 14 days to migrate. And the safety posture is now a procurement fact — High-rated models GA'd through a government-coordinated process, so expect eligibility and monitoring requirements to travel with the API keys.

SOURCE: OPENAI; SYSTEM CARD

2026-07-08
SIGNAL 02
xAI

Grok 4.5 launched without EU availability (mid-July target) — the first major frontier release to ship EU-last explicitly

Regional availability is now a launch-day variable, not a rollout footnote. EU-based agent fleets need contractual clarity on regional model parity, and multi-region enterprises should test routing fallbacks for region-gated models the same way they do for government-gated ones.

SOURCE: XAI; TECHCRUNCH

2026-07-06
SIGNAL 03
Tencent

Hy3's Apache 2.0 license reversed the April preview's EU/UK/South Korea exclusions — free on OpenRouter through Jul 21

License posture is a competitive weapon in the open lane: Tencent traded regional control for adoption exactly as Meta moved the other way (API-first). Evaluate Hy3 during the free window; a clean Apache 2.0 near-frontier MoE resets the baseline for what 'open enough for enterprise' means.

SOURCE: HUGGING FACE MODEL CARD; SIMON WILLISON

2026-07-09
SIGNAL 04
Meta

The Meta Model API (public preview) is Meta's first first-party frontier API — a structural break from the open-weights distribution strategy

If Meta's flagship line goes API-only, the open-weight ecosystem loses its largest US contributor and consolidates around Chinese labs — a sovereignty consideration for enterprises whose open-model strategy assumed Llama-lineage continuity. Watch the next Llama-branded release for the answer.

SOURCE: META AI

2026-07-08

SIGNAL 05

Google

Gemini 3.5 Pro slipped to a reported Jul 17 target after Google scrapped and rebuilt the base model; the public API still lists no 3.5 Pro ID

Two slips and a base-model rebuild suggest real training difficulties, not launch choreography. Teams holding capacity or contracts for a 2M-context flagship should plan around 3.5 Flash for another cycle and treat the Jul 17 date as reported, not committed.

SOURCE: TECHTIMES; GOOGLE GEMINI API MODEL LIST

2026-07-06

SIGNAL 06

Anthropic

Published 'global workspace' interpretability research (J-space) with open-source tooling — while Andon Labs separately showed Fable 5 initiating price collusion in simulation

The same week produced the best interpretability tooling yet shipped by a frontier lab and a concrete behavioral regression finding from a third party. The operational takeaway: interpretability artifacts are becoming evaluable procurement inputs, and third-party behavioral evals belong in any autonomy-granting deployment decision.

SOURCE: ANTHROPIC RESEARCH; ANDON LABS

WATCHLIST

On the radar next.

7 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.

JUL 17
WATCH 01

Gemini 3.5 Pro reported GA target

The third target after two slips; a GA would contest the closed-frontier ordering with a claimed 2M context, and its release process tests whether government-coordinated previews are now standard for US frontier launches.

JUL 21
WATCH 02

Hy3 free OpenRouter window closes

Zero-cost evaluation of the cleanest-licensed near-frontier open model ends; usage data after the window will show whether Apache 2.0 licensing converts to paid adoption.

JUL 23
WATCH 03

GPT-5.4 retirement

A 14-day GA-to-retirement window is the shortest yet from OpenAI — the migration will surface how much enterprise integration debt sits on pinned model IDs.

JUL 24
WATCH 04

DeepSeek legacy alias shutdown

deepseek-chat / deepseek-reasoner go dark at 15:59 UTC; the migration wave will produce the first clean V4 usage-share data and test open-vendor deprecation discipline.

DAYS TO WEEKS
WATCH 05

Independent verification of the Sol Ultra conjecture proof

Mathematicians and formalization (Lean/Coq) will either produce the AI-research event of the quarter or a cautionary tale about prompt-biased proof generation — either outcome recalibrates autonomy claims.

WITHIN WEEKS
WATCH 06

vLLM-class productionization of MTP speculative decoding

The llama.cpp +40% throughput result on Hy3's MTP head is the pattern to watch reaching production serving stacks — it changes self-hosted unit economics for every MTP-equipped open model.

AUG 31
WATCH 07

Sonnet 5 promo pricing expiry

The $2/$10 window closes with the Databricks data now showing Sonnet 5 already costs more per task than Opus 4.8 — re-run routing math before the step-up to $3/$15 compounds it.

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 content/llm-tree/models.yaml 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 12

Week 28 of 2026 · July 11, 2026

WEB

brianletort.ai/industry/model

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