THE MODEL PULSE ISSUE 09 · WEEK 25 OF 2026

THE BIG READ

Open weights took the lead while the closed frontier stalled — and the benchmark itself was rewritten.

KEY SIGNALS THIS PERIOD

2	4	3	2
Models added to the tree	Vendors shipped frontier-class	Architectural patterns crossed multiple vendors	Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

June 20, 2026

Issue 09 · 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.

W25 was a loud week for open weights and a quiet one at the closed frontier. Z.ai shipped GLM-5.2 under a genuine MIT license — a ~744B / ~40B-active sparse-attention MoE with 1M context that independent testing (VentureBeat) says beats GPT-5.5 on several long-horizon coding benchmarks at roughly one-sixth the cost, and that Artificial Analysis now cites as the leading open-weight model. MiniMax-M3's sparse-attention weights matured in-window with an arXiv report validating its efficiency claims, though its non-OSI Community License gates commercial use. The closed frontier, by contrast, marked time: no GA from OpenAI (GPT-5.6 remains rumor) or xAI, Gemini 3.5 Pro slipped from June to July, and Anthropic's Claude Fable 5 stayed government-suspended the entire week (Opus 4.8 is the working leader). The third shift was measurement itself — Artificial Analysis rebased its Intelligence Index to v4.1, re-weighting the industry's headline benchmark around agentic tasks, so scores are no longer back-comparable to v4.0. The procurement implication: open self-host is now a live coding option, not a hedge; teams should pilot MIT-licensed GLM-5.2, read every 'open' license carefully (GLM-5.2 MIT vs MiniMax Community), and re-baseline evaluations on the agentic v4.1 index while keeping closed-frontier fallbacks given the demonstrated availability risk.

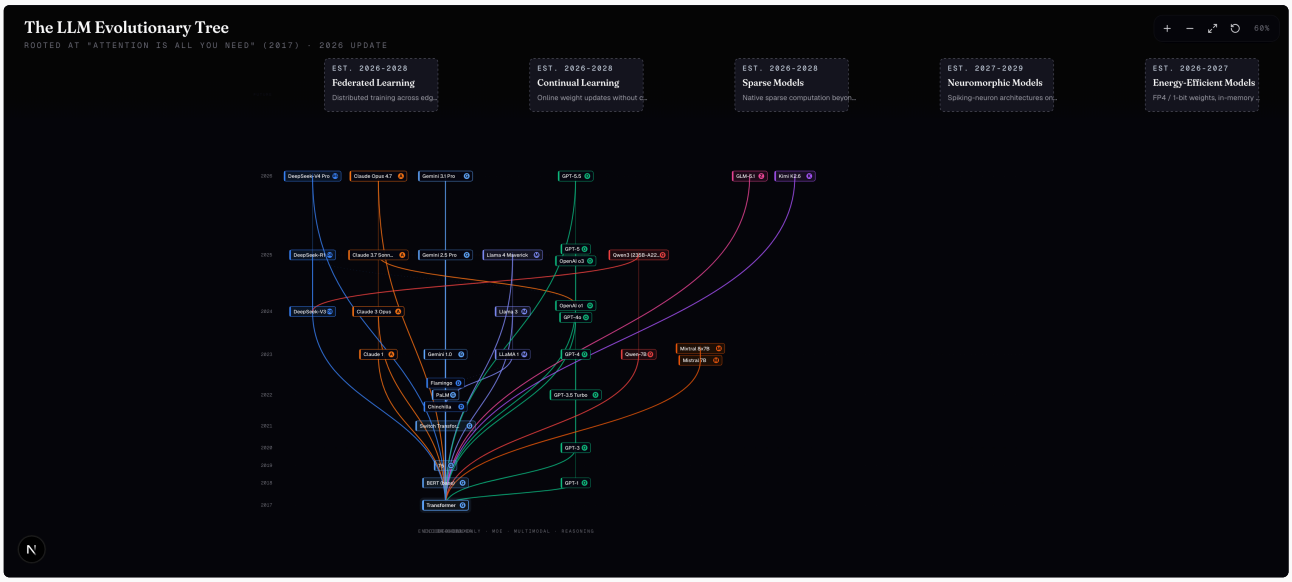
THE BAR

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.

Two W25 additions: GLM-5.2 (MIT open-weight frontier-adjacent MoE that took the open lead) and MiniMax-M3 (sparse-attention multimodal MoE, weights matured with arXiv verification).



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

ADDED (2)

glm-5-2
minimax-m3

UPDATED (0)

No updates to existing rows this period.

No new closed frontier model entered the tree: GPT-5.6 is rumor only, Gemini 3.5 Pro slipped to July, and Claude Fable 5 (added W24) stayed suspended. ByteDance's seed-2.1-pro-preview is excluded as an undisclosed preview.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-06-09

RELEASE 01

Anthropic

FRONTIER

REASONING

Claude Fable 5 (still suspended)

Remained government-suspended for the full week; #1 on the rebased Artificial Analysis Index (60) but unavailable, so Opus 4.8 (56) is the top available closed model

The closed-frontier leader on paper is unusable in practice for a second week, which keeps the availability/sovereign risk live. Architects should treat the AA top score as aspirational and standardize on the top available model (Opus 4.8) with fallback routing, not on a suspended SKU.

SOURCE: ANTHROPIC; ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V4.1

2026-07 TARGET

RELEASE 02

Google DeepMind

FRONTIER

REASONING

Gemini 3.5 Pro (slipped to July)

GA slipped from June to July: still a limited Vertex preview with no public model card, pricing, or independent benchmark vs Fable 5/Opus 4.8

A frontier movement by absence for the third consecutive issue. Buyers should keep an evaluation slot ready but not pause current baselines; the closed frontier's cadence is visibly slipping while open weights accelerate.

SOURCE: BUSINESS INSIDER; GOOGLE

OPEN WEIGHTS

Open-frontier and open-source drops.

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

2026-06-16

RELEASE 01

Z.ai

OPEN FRONTIER

MOE

GLM-5.2

MIT-licensed ~744B / ~40B-active sparse-attention MoE with 1M context; independent testing says it beats GPT-5.5 on long-horizon coding at ~1/6 the cost and Artificial Analysis cites it as the top open-weight model

GLM-5.2 is the week's most consequential release: a truly permissive (MIT, no regional limits) frontier-adjacent model that is self-hostable and sovereignty-friendly for long-context agentic coding. Architects should pilot it for self-host/private coding workloads and use its ~$1.40/$4.40 per-MTok API as a pricing benchmark against closed flagships.

SOURCE: Z.AI BLOG; VENTUREBEAT; HUGGING FACE

2026-06-12

RELEASE 02

MiniMax

OPEN FRONTIER

MOE

MiniMax-M3

428B / ~23B-active sparse-attention MoE with native multimodality and 1M context; weights matured with an arXiv report verifying ~9x/15x prefill/decode efficiency, under a non-OSI Community License

MiniMax-M3 combines frontier-adjacent coding, genuine 1M context, and native multimodality in one downloadable checkpoint — but the MiniMax Community License gates commercial use, so 'open weights' here does not mean free to deploy. Teams should verify the efficiency claims via the arXiv report and clear the license before planning a commercial deployment.

SOURCE: TECHTIMES; HUGGING FACE; ARXIV:2606.13392

ARCHITECTURE WATCH

Patterns to track.

3 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

Open weights close the cost gap

GLM-5.2 (MIT) | MiniMax-M3 | Grok 4.3 on Bedrock

Frontier-adjacent capability is collapsing toward commodity inference pricing. GLM-5.2's MIT weights reportedly match or beat GPT-5.5 on long-horizon coding at ~1/6 the cost, and API list prices (GLM-5.2 ~$1.40/$4.40 per MTok; Grok 4.3 $1.25/$2.50 on Bedrock) keep falling. Procurement should pilot open self-host for routine and long-context coding and reserve closed flagships for the highest-risk reasoning.

SOURCE: Z.AI; VENTUREBEAT; AMAZON BEDROCK

PATTERN 02

The headline benchmark pivots to agents

Artificial Analysis Intelligence Index v4.1 | LMArena Agent Arena

Artificial Analysis rebased its Intelligence Index to v4.1, re-weighting around agentic tasks (GDPval-AA v2 at 20%, Terminal-Bench, banking agents) and dropping a saturated benchmark, while LMArena's Agent Arena scores behavioral signals (retries, steerability) rather than preference votes. Boards comparing models on 'the AA Index' must note v4.1 scores are not back-comparable to v4.0; re-baseline evaluation harnesses now.

SOURCE: ARTIFICIAL ANALYSIS; ARENA.AI CHANGELOG

PATTERN 03

License divergence within 'open'

GLM-5.2 (MIT) | MiniMax-M3 (Community License)

Two of the week's open releases sit on opposite ends of the permissiveness spectrum: GLM-5.2 under MIT with no regional limits versus MiniMax-M3 under a Community License that gates commercial use. For enterprise adoption, 'open weights' does not equal 'free to deploy commercially' — legal and procurement should read the actual license before standardizing on a model.

SOURCE: Z.AI; MINIMAX HUGGING FACE CARD

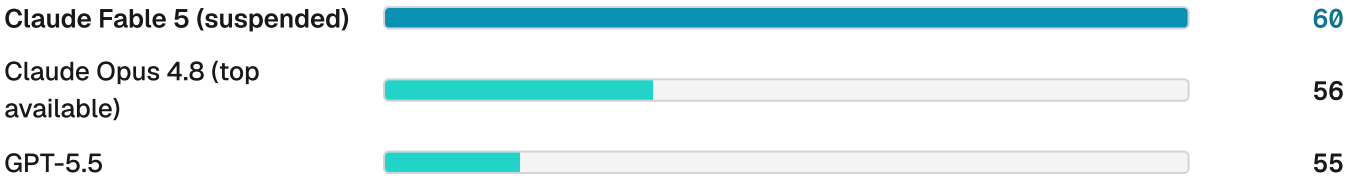
BENCHMARK MOVES

Where the leaderboard moved.

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

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V4.1 MOVE 01

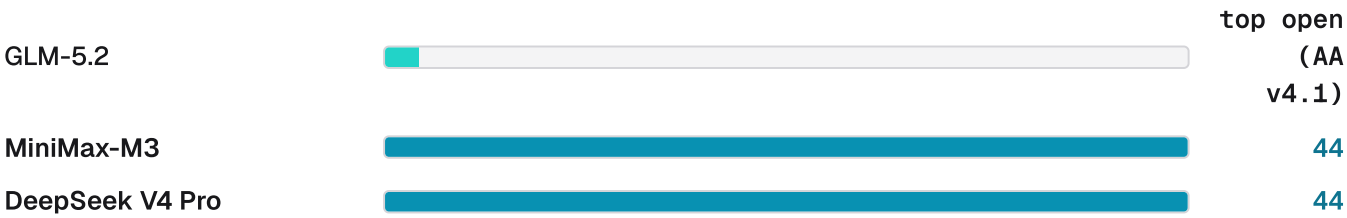
Methodology rebased around agentic tasks (Jun 15); leaders Fable 5 = 60 (top but suspended), Opus 4.8 = 56 (top available), GPT-5.5 = 55; scores not back-comparable to v4.0



SOURCE: ARTIFICIAL ANALYSIS

OPEN-WEIGHT LEADERBOARD MOVE 02

GLM-5.2 took the open-weight lead in-window; MiniMax-M3 and DeepSeek V4 Pro sit at ~44 on the rebased index



SOURCE: ARTIFICIAL ANALYSIS; VENTUREBEAT

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Jun 20, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Opus 4.8**	GPT-5.5	Fable 5 leads the AA v4.1 index (60) but stayed suspended all week; Opus 4.8 (56) is the top available closed model.
Open frontier	**GLM-5.2**	MiniMax-M3	GLM-5.2's MIT release took the open-weight lead in-window; MiniMax-M3 contends but is gated by a non-OSI license.
Reasoning	**Claude Opus 4.8**	GPT-5.5	Closed reasoning leadership steady among available models while Gemini 3.5 Pro slipped to July.
Coding	**Claude Opus 4.8**	GLM-5.2	Open weights are closing fast on long-horizon coding; GLM-5.2 reportedly beats GPT-5.5 at ~1/6 the cost.
Multimodal	**Gemini 3.1 Pro**	MiniMax-M3	Gemini remains the general multimodal reference; MiniMax-M3 adds native-multimodal open weights.
Edge / small	**Mellum2**	North Mini Code	Efficient open coding/sub-agent models unchanged in-window; the week's open action was at the frontier-adjacent tier.

VENDOR SIGNALS

Pricing, gating, deprecation.

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

2026-06-16
SIGNAL 01
Z.ai

Released GLM-5.2 under MIT with API pricing ~$1.40/$4.40 per MTok (~1/6 of comparable frontier)

A tier-1 permissive open release at commodity pricing pressures every closed flagship's price/value story. Procurement should use GLM-5.2 as a negotiating anchor and pilot it for self-host coding; investors should treat open-weight pricing as a structural deflationary force on inference.

SOURCE: Z.AI; DATANORTH

2026-06-15
SIGNAL 02
Artificial Analysis

Rebased the Intelligence Index to v4.1, re-weighting around agentic workloads; v4.1 scores are not back-comparable to v4.0

The industry's headline benchmark now measures agentic capability, not static Q&A. Boards and architects must re-baseline model comparisons on v4.1 and avoid mixing old and new index numbers in procurement decisions.

SOURCE: ARTIFICIAL ANALYSIS

2026-06-15
SIGNAL 03
xAI

Grok 4.3 went GA on Amazon Bedrock ($1.25/$2.50 per MTok, 1M context), making xAI the third independent frontier lab on Bedrock alongside Anthropic and OpenAI

CIOs can now evaluate all three independent US frontier labs under one IAM and billing surface. The caveat is a non-standard endpoint and a context-window pricing cliff above 200K tokens; this is distribution, not a new capability tier.

SOURCE: DIGITALAPPLIED; MEMEBURN

2026-06-15
SIGNAL 04
Anthropic

Claude Fable 5 and Mythos 5 remained government-suspended all week; the planned Jun 23 usage-credit subscription change is moot while access is off

The top-tier closed model's availability is still a sovereign/regulatory variable, not an SLA. Buyers should keep Opus 4.8/Sonnet fallbacks wired and avoid single-sourcing the frontier for production-critical paths.

SOURCE: ANTHROPIC

WATCHLIST

On the radar next.

4 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.

JULY
WATCH 01

Gemini 3.5 Pro GA

Pro slipped to July. Its GA and first independent AA v4.1 pass will show whether Google can re-take a frontier lead now contested by both Opus 4.8 and a surging open-weight field.

JUN-AUG
WATCH 02

Claude Fable 5 / Mythos 5 restoration

Restoration terms (geo-gating, KYC, or a permanent civilian/government capability split) will set the precedent for sovereign access risk and decide whether Fable 5 re-enters the available scorecard.

JUN-AUG
WATCH 03

GLM-5.2 adoption and independent SWE-Bench replication

Downloads, integrations, and third-party benchmark replication will show whether MIT-licensed open weights become production substrate and force closed-flagship price cuts.

JUN-JUL
WATCH 04

GPT-5.6 / next OpenAI flagship

Codenames and prediction markets pointed to a launch just after this window. A real system card would re-set the closed frontier and test whether OpenAI answers the open-weight cost pressure.

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 09

Week 25 of 2026 · June 20, 2026

WEB

brianletort.ai/industry/model

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