THE MODEL PULSE ISSUE 08 · WEEK 24 OF 2026

THE BIG READ

A new closed frontier shipped and was switched off in the same week; the durable open progress was a diffusion efficiency path.

KEY SIGNALS THIS PERIOD

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

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

June 13, 2026

Issue 08 · 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.

W24's model story had two halves. Anthropic shipped Claude Fable 5 — a new Mythos-class tier above Opus — and it debuted #1 on the independent Artificial Analysis Intelligence Index at 64.9, roughly five points ahead of GPT-5.5. Three days later (Jun 12), a U.S. government export-control directive forced Anthropic to disable Fable 5 and its safeguards-lifted Mythos 5 sibling for every customer, routing queries back to Opus 4.8. So the public closed-frontier leader for half a week is, at week's end, unavailable — the working default is again Opus 4.8. Meanwhile the open layer's real progress was efficiency, not a new crown: Google DeepMind released DiffusionGemma, an Apache-2.0 text-diffusion model that generates ~1,000+ tok/s on an H100, and Cohere shipped North Mini Code, a cheap self-host coding MoE; MiniMax-M3 was announced (AA Index 55) but its weights are still pending. Gemini 3.5 Pro remained not-GA. The procurement implication is sharper than another leaderboard reshuffle: a top-tier closed model can now be revoked by a third party, so production systems need a standardized eval harness and a hard fallback router, with cheap open/local models carrying routine work and the closed frontier reserved for high-value reasoning.

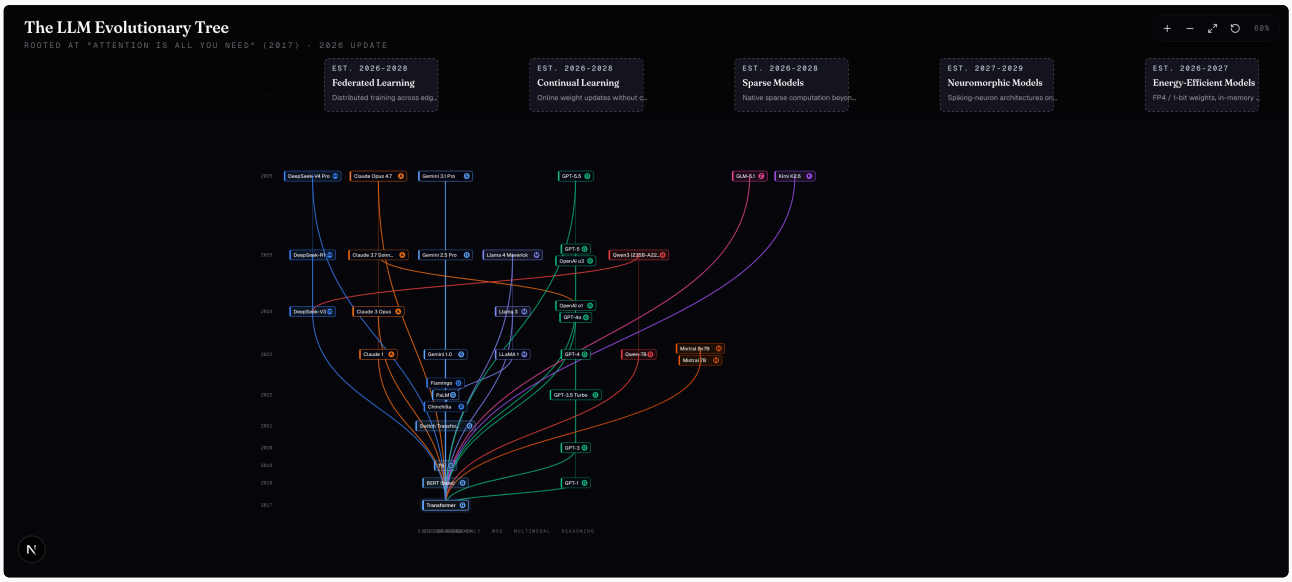
THE BAR

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 W24 additions: Claude Fable 5 (new Mythos-class closed frontier), DiffusionGemma (open text-diffusion), and Cohere North Mini Code (efficient open coding MoE).



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

ADDED (3)

- `claude-fable-5`
- `diffusiongemma`
- `north-mini-code`

UPDATED (0)

No updates to existing rows this period.

Mythos 5 is the same model as Fable 5 with cyber safeguards lifted (government-only) and is not a separate tree row. MiniMax-M3 is excluded until its weights are released, and Gemini 3.5 Pro is excluded until Google publishes a GA model card or API identifier.

FRONTIER MOVEMENTS

Flagship-class releases.

3 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

New Mythos-class tier above Opus; debuted #1 on the Artificial Analysis Intelligence Index at 64.9, priced $10/$50 per MTok

Fable 5 is the new public top of the closed frontier on independent indices, but it is priced as an async heavy-lift tool (~2x Opus 4.8) and was free on paid tiers only through Jun 22 before reverting to usage credits. Architects should reserve it for high-value reasoning and codebase-scale orchestration, not default every agent call to it, and should re-benchmark on cost-per-task rather than headline scores.

SOURCE: ANTHROPIC LAUNCH; ARTIFICIAL ANALYSIS INTELLIGENCE INDEX

2026-06-12

RELEASE 02

Anthropic

FRONTIER

REASONING

Claude Mythos 5 (and the Jun 12 takedown)

Same model as Fable 5 with safeguards lifted for a government program; both force-disabled globally by a U.S. export-control directive three days after launch

This is the first known government-forced takedown of a deployed frontier model, and it reframes model selection: a top-tier closed model is now a sovereign/regulatory single point of failure that a third party can switch off. Architects and operators should build hard fallbacks (queries already revert to Opus 4.8) and avoid single-sourcing the frontier for any production-critical path.

SOURCE: VENTUREBEAT; MARKTECHPOST (REUTERS/BBC/AXIOS CITED SECONDHAND)

2026-06 TARGET

RELEASE 03

Google DeepMind

FRONTIER

REASONING

Gemini 3.5 Pro

Still pending at W24 close: limited Vertex enterprise preview, no public API identifier, price row, or independent benchmark despite a Google-stated June target

A frontier movement by absence for the second consecutive issue. Buyers should keep the launch on the June watchlist but not pause current baselines on an unreleased SKU; the likely enterprise pattern remains task routing (Gemini for huge-context/multimodal, Anthropic/OpenAI for coding and agentic reliability) rather than a universal replacement.

SOURCE: TECHTIMES; PRESENC LAUNCH BRIEF

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.

2026-06-10

RELEASE 01

Google DeepMind

SPECIALIST MOE

DiffusionGemma

Apache-2.0 open text-diffusion model: 26B / ~3.8B-active MoE on a Gemma 4 backbone, multimodal-in, 256K context, ~1,000+ tok/s on an H100 with native vLLM

DiffusionGemma is the week's most architecturally significant open release: a credible open path to low-latency, local interactive inference via block diffusion rather than autoregressive decoding. Teams building latency-sensitive local agents should evaluate it as an efficiency option and watch whether the autoregressive-vs-diffusion split becomes a real procurement axis.

SOURCE: GOOGLE BLOG; HUGGING FACE MODEL CARD

2026-06-09

RELEASE 02

Cohere

EDGE / SMALL

MOE

North Mini Code

Small open-weight coding MoE (~30B / ~3B active) scoring 27.6 on the Artificial Analysis Intelligence Index

North Mini Code is not frontier-competitive, but it is a cheap self-host candidate for routine code and text sub-agent calls. Teams trying to cut orchestration cost should treat it as a routing node for low-risk repeated work, reserving closed flagships for high-stakes reasoning.

SOURCE: ARTIFICIAL ANALYSIS

2026-06-08

RELEASE 03

MiniMax

OPEN FRONTIER

MOE

MiniMax-M3 (announced; weights pending)

Announced at 55 on the Artificial Analysis Intelligence Index — would be the leading open-weights model once weights release, but is not yet downloadable

MiniMax-M3 is the strongest competitively-scoring open item of the window, but with weights pending it is a signal, not a deployable asset. Procurement teams tracking an open-frontier alternative should prepare an evaluation slot but not plan a migration until the weights and an independent benchmark land.

SOURCE: ARTIFICIAL ANALYSIS

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

Government as a model kill switch

`Claude Fable 5 takedown` `Claude Mythos 5 (Project Glasswing)`

The Jun 12 export-control directive that forced Anthropic to disable Fable 5 and Mythos 5 globally turns regulatory access into an architecture variable. A model that is best-in-class on Monday can be unavailable by Friday for reasons outside the vendor's control. Enterprise architecture should now treat top-tier closed models as revocable dependencies and design fallback routing and capability-degradation plans accordingly.

SOURCE: VENTUREBEAT; MARKTECHPOST

PATTERN 02

Text diffusion enters open weights

`DiffusionGemma` `Gemma 4 backbone`

DiffusionGemma brings block-diffusion text generation into the open-weight mainstream with native vLLM support and ~1,000+ tok/s throughput. The pattern matters because it offers a different latency/throughput profile than autoregressive decoding for interactive and local workloads. Architects evaluating local inference should add a diffusion option to their benchmark set rather than assuming autoregressive is the only path.

SOURCE: GOOGLE BLOG; HUGGING FACE

PATTERN 03

Frontier shipping concentrated in one vendor

`Anthropic Fable 5` `Gemini 3.5 Pro pending` `GPT-5.6 rumor only`

For the second straight issue, the only new frontier model came from Anthropic, while Google's Pro slipped and OpenAI's next model stayed a rumor. That concentration means the frontier's cadence — and now its availability — depends heavily on a single lab. Buyers should diversify model routing across vendors and keep a credible open/local tier so a single vendor's release or takedown cannot stall production.

SOURCE: ANTHROPIC LAUNCH; ARTIFICIAL ANALYSIS; TECHTIMES

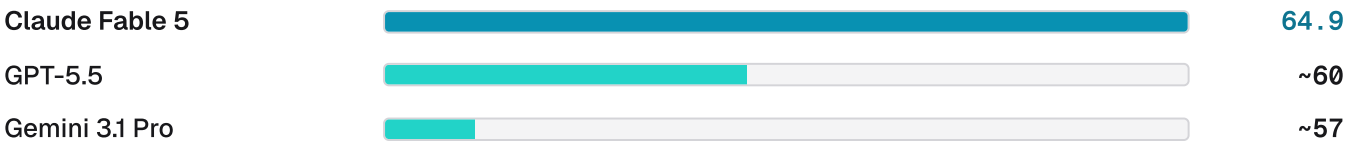
BENCHMARK MOVES

Where the leaderboard moved.

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

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX MOVE 01

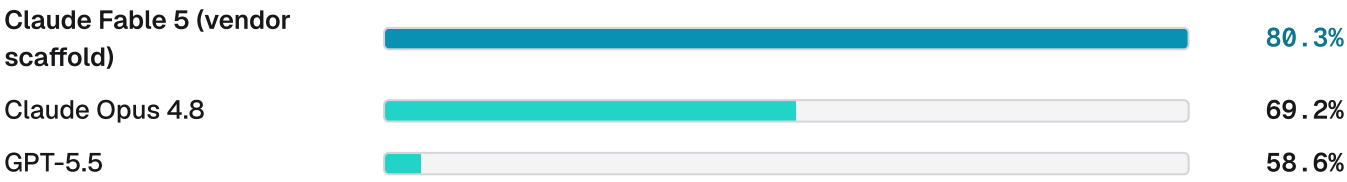
Claude Fable 5 debuted #1 at 64.9, roughly five points ahead of GPT-5.5, and set the highest score on five of the ten underlying benchmarks



SOURCE: ARTIFICIAL ANALYSIS

SWE-BENCH PRO (CODING) MOVE 02

Fable 5's vendor-scaffold 80.3% widens the headline closed lead, but it is unreplicated by Scale SEAL/Epoch; independent vals.ai shows 95% on SWE-Bench Verified



SOURCE: MORPHLLM / VALS.AI; VENDOR HARNESS (CONTESTED)

OPEN-WEIGHT INTELLIGENCE MOVE 03

MiniMax-M3 (55, weights pending) would lead open weights once released; North Mini Code lands at 27.6 in the efficient tier

SOURCE: 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 Jun 13, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	Claude Opus 4.8	Claude Fable 5 (suspended)	Fable 5 debuted #1 (AA 64.9) but was force-disabled Jun 12; Opus 4.8 is the working default again until access is restored.
Open frontier	DeepSeek V4-Pro	MiniMax-M3 (weights pending)	No open frontier text model displaced the April leaders in-window; MiniMax-M3's 55 AA Index would contend once weights ship.
Reasoning	Claude Opus 4.8	GPT-5.5	Fable 5 set the reasoning bar before its takedown; with it suspended, Opus 4.8 leads the available closed reasoning tier.
Coding	Claude Opus 4.8	GPT-5.5	Fable 5's 80.3% SWE-Bench Pro is a contested vendor figure and the model is suspended; Opus 4.8 remains the dependable coding default.
Multimodal	Gemini 3.1 Pro	DiffusionGemma	Gemini remains the general multimodal reference; DiffusionGemma opens a low-latency open diffusion branch.
Edge / small	Mellum2	North Mini Code	Efficient open coding/sub-agent models keep widening; North Mini Code joins Mellum2 as a cheap self-host routing node.

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.

2026-06-12
SIGNAL 01
Anthropic

A U.S. export-control directive forced Anthropic to disable Fable 5 and Mythos 5 for all customers; queries fall back to Opus 4.8

Top-tier model availability is now a sovereign/regulatory variable, not just an uptime SLA. Buyers should require contractual clarity on access continuity and design fallback routing before standardizing on any single closed flagship.

SOURCE: VENTUREBEAT; MARKTECHPOST

2026-06-09
SIGNAL 02
Anthropic

Fable 5 priced at $10/$50 per MTok (~2x Opus 4.8); included on paid tiers only through Jun 22, then reverting to usage credits

The top-tier price ceiling moved up and the free-access window is short. Operators should model the post-Jun 22 cost cliff and route only high-value work to Fable 5, keeping cheaper models for routine calls.

SOURCE: ANTHROPIC LAUNCH

2026-06-11
SIGNAL 03
xAI

Launched the Grok Build Plugin Marketplace bundling skills, MCP servers, and agents (commit-SHA-pinned), with partners MongoDB, Vercel, Sentry, Chrome DevTools, and Cloudflare

xAI is competing on agent tooling and distribution rather than a new frontier model this window. Platform teams should weigh marketplace lock-in and pinning guarantees when evaluating agent ecosystems.

SOURCE: X.AI

2026-06-11
SIGNAL 04
OpenAI

Acquired Ona (Gitpod) for cloud sandboxes that keep long-running agents alive, folding it into Codex (5M+ weekly users)

Durable long-running-agent infrastructure is being absorbed into the model layer, narrowing 'agent persistence' as a standalone wedge. Architects building long-horizon agents should expect this capability to become a platform feature rather than a separate purchase.

SOURCE: SILICONANGLE

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.

JUN 14-30

WATCH 01

Fable 5 / Mythos 5 access restoration

Whether Anthropic restores a top-tier model — and with what geo-gating or KYC controls — sets the precedent for sovereign access risk and determines whether Fable 5 re-enters the closed-frontier scorecard.

JUN 14-30

WATCH 02

Gemini 3.5 Pro GA and first independent benchmark

Google's Pro has slipped to the edge of its June commitment. A GA with an independent Artificial Analysis pass will show whether it lands above or below Fable 5 and GPT-5.5.

JUN-JUL

WATCH 03

MiniMax-M3 weights release

If MiniMax ships the weights at its announced 55 AA Index, it becomes the leading open-weights model and a credible self-host alternative to closed frontiers.

JUN-AUG

WATCH 04

Independent SWE-Bench Pro replication of Fable 5

Fable 5's 80.3% is a vendor-harness figure. An Epoch or Scale SEAL replication will confirm or deflate the coding-lead claim that buyers are being asked to act on.

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 08

Week 24 of 2026 · June 13, 2026

WEB

brianletort.ai/industry/model

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