THE MODEL PULSE
ISSUE 11 · WEEK 27 OF 2026

THE BIG READ

Washington entered the release pipeline: GPT-5.6 launched gated to ~20 approved partners, Fable 5 came back, and Sonnet 5 split sticker price from cost-per-task.

KEY SIGNALS THIS PERIOD

5	5	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 4, 2026
Issue 11 · 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.

Within five days the US government both gated a new frontier launch and un-gated a suspended one. OpenAI previewed the GPT-5.6 family (Sol, Terra, Luna) on June 26 to roughly 20 government-approved partner organizations — the first US frontier model launched under a government-managed access list — while Commerce withdrew the Anthropic export-control order on June 30 and Claude Fable 5 returned globally on July 1. The procurement implication is direct: frontier availability is now partly a regulatory variable, so contracts and architectures must assume any flagship can be gated at launch or suspended after it, and multi-model routing is availability insurance, not just cost optimization. The week's biggest GA release, Claude Sonnet 5 (June 30, 1M context, 85.2% SWE-bench Verified, default for Claude Free/Pro), delivered the second structural lesson: at max effort its token appetite pushes measured cost to $2.29 per task — ~15% above Opus 4.8 — despite a far lower per-token price, so sticker price and cost-per-task have formally diverged. Buyers should route by measured cost-per-completed-task with effort level as an explicit parameter, and re-run that math before Sonnet 5's introductory pricing lapses on August 31. Meituan's MIT-licensed LongCat-2.0 rounds out the week by adding demand-proven open-weight pressure — and a claimed NVIDIA-free trillion-scale training run — to the cost lane.

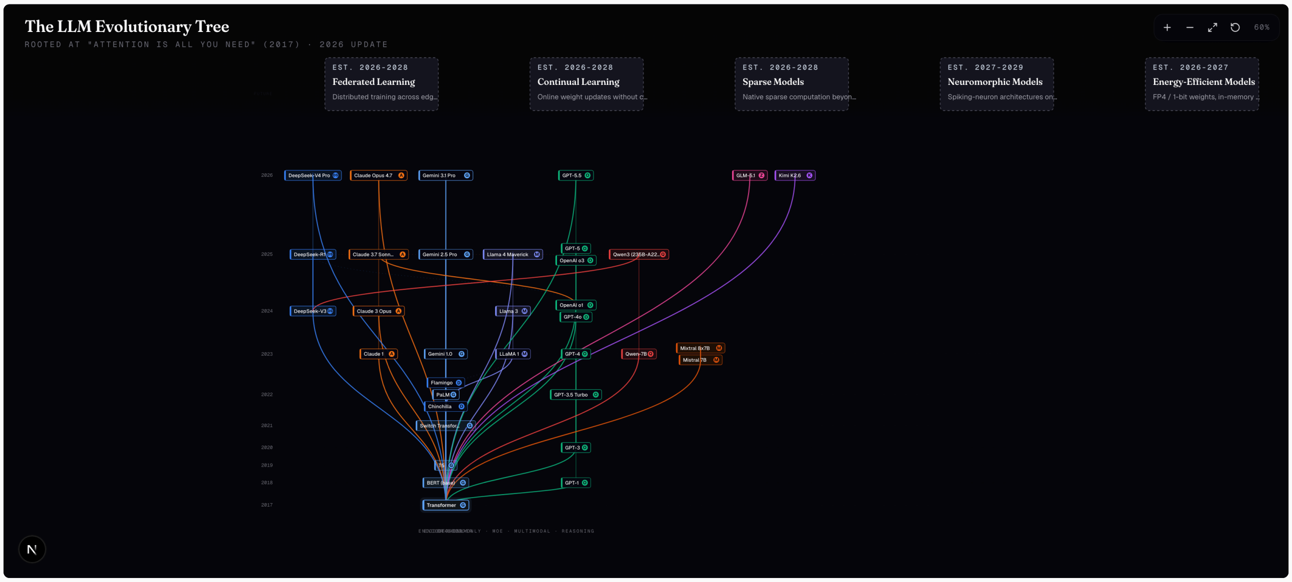
THE BAR

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: claude-sonnet-5 and gpt-5-6-sol extend the closed reasoning canopy, longcat-2-0 lands a 1.6T open MoE, leanstral-1-5 opens a formal-verification specialist node, and nemotron-labs-twotower extends the diffusion-retrofit lineage. One row updated: claude-fable-5 now records the July 1 restoration.



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

ADDED (5)

- claude-sonnet-5
- gpt-5-6-sol
- longcat-2-0
- leanstral-1-5
- nemotron-labs-twotower

UPDATED (1)

- claude-fable-5

gpt-5-6-sol carries status gated (preview limited to ~20 government-approved partners, no independent benchmarks). The unverified Pulpie family (Feyn Inc.) was excluded as below the significance and sourcing threshold.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-06-30
RELEASE 01
Anthropic
REASONING
AGENTIC

Claude Sonnet 5

The week's biggest GA release: 1M context, 85.2% SWE-bench Verified, effort dials, at introductory $2/$10 per M tokens

Sonnet 5 gives architects near-Opus agentic quality at a mid-tier sticker price — independent testing shows it beating Opus 4.8 on agentic knowledge-work benchmarks — but at max effort it costs ~15% more per completed task than Opus 4.8. Route by measured cost-per-task, not per-token price, and model post-promo economics (standard $3/$15 from September 1) before committing volume.

SOURCE: ANTHROPIC; ARTIFICIAL ANALYSIS

2026-06-26
RELEASE 02
OpenAI
SPECIALIST
REASONING

GPT-5.6 family (Sol / Terra / Luna)

Next OpenAI flagship generation previewed — but launched to only ~20 government-approved partner organizations

Marked specialist because it is gated: this is the first US frontier launch under a government-managed access list, so procurement teams should treat release gating as a live availability risk for every future frontier upgrade and keep multi-vendor fallbacks tested. Terra's claimed GPT-5.5-level capability at half the price is the pricing lever to model ahead of GA.

SOURCE: OPENAI

2026-07-01
RELEASE 03
Anthropic
FRONTIER
REASONING

Claude Fable 5

Restored globally July 1 after Commerce withdrew the June 12 export-control order; back atop the available boards

The strongest model on the public leaderboards is purchasable again, resetting the practical-leader question W26 settled in Opus 4.8's favor — but the 19-day forced outage is now a demonstrated failure mode, so any single-model dependency on a frontier flagship needs a tested fallback path. Note the fine print: after July 7, Fable 5 moves to usage credits even on paid plans, so treat it as a premium metered resource in budgets.

SOURCE: ANTHROPIC; VENTUREBEAT

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.

2026-06-30

RELEASE 01

Meituan

OPEN FRONTIER

MOE

LongCat-2.0

OpenRouter's stealth chart-leader 'Owl Alpha' unmasked: a 1.6T MoE with native 1M context under MIT, trained on 50K+ domestic Chinese ASICs

A near-frontier agentic-coding MoE with two months of real developer demand joins GLM-5.2 and DeepSeek V4 in the open cost-pressure lane — benchmark it wherever unit economics beat absolute frontier quality, but verify actual weight availability before treating it as self-hostable. The claimed NVIDIA-free 35T-token training run, if it holds, weakens the assumption that export controls throttle Chinese frontier training; factor that into sovereignty and supply planning.

SOURCE: HUGGING FACE MODEL CARD; VENTUREBEAT; SILICONANGLE

2026-06-30

RELEASE 02

Mistral AI

SPECIALIST MOE

Leanstral 1.5

Apache 2.0 Lean 4 proof-engineering agent (119B MoE / 6.5B active) saturates miniF2F and solves 587/672 PutnamBench problems at ~$4 each

The strongest evidence yet that narrow verticalized agent models can beat frontier brute force on economics — ~75x cheaper per problem than Seed-Prover 1.5 high — with a practical hook for engineering leaders: Mistral reports 5 previously unknown bugs found across 57 open-source repositories via formal code verification. Self-host the Apache weights for anything durable; the free labs API endpoint retires September 30.

SOURCE: MISTRAL AI; MARKTECHPOST

2026-07-01

RELEASE 03

NVIDIA

SPECIALIST MOE

Nemotron-Labs-TwoTower

Open block-diffusion model on a frozen AR backbone: 2.42x wall-clock throughput at 98.7% of autoregressive quality

TwoTower shows diffusion decoding can be retrofitted onto an existing pretrained AR checkpoint for ~2.1T training tokens instead of a 25T re-pretrain — a cheap serving-throughput upgrade path any lab with a strong AR model can copy, which compounds W26's serving-economics story. One checkpoint supports diffusion, mock-AR, and standard AR decoding, so platform teams can A/B the pattern with low switching risk once vLLM-class support lands.

SOURCE: ARXIV 2606.26493; MARKTECHPOST

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

Effort dials decouple sticker price from cost-per-task

`Claude Sonnet 5` `DeepSeek V4` `GPT-5.5`

Every major vendor now ships user-controllable reasoning-effort settings, and this week produced the clearest data point yet that token-hungry effort scaling inverts price intuition: Sonnet 5 at max effort costs $2.29 per Intelligence Index task at standard pricing — more per completed task than the nominally pricier Opus 4.8. Evaluation harnesses must price the task, not the token, and effort level must become an explicit routing parameter in every model gateway.

SOURCE: ARTIFICIAL ANALYSIS

PATTERN 02

Diffusion decoding retrofitted onto frozen AR backbones

`Nemotron-Labs-TwoTower` `DiffusionGemma`

Two vendors in a month have shipped open diffusion LLMs built on top of existing autoregressive checkpoints rather than trained from scratch — TwoTower needed only ~2.1T training tokens against its backbone's 25T. Diffusion is emerging as a cheap post-hoc throughput upgrade (2.42x wall-clock at 98.7% quality) rather than a rival pretraining paradigm; platform teams should watch for this reaching production serving stacks via vLLM-class support.

SOURCE: ARXIV 2606.26493

PATTERN 03

1M-context sparse attention is the new open-weight table stakes

`DeepSeek V4` `LongCat-2.0` `GLM-5.2`

The Chinese open-weight labs have converged on the same recipe: sparse/compressed attention variants that make 1M-token context economically routine. The DeepSeek V4 technical report published this window puts hard numbers on it — roughly 90% KV-cache reduction and 27% of V3.2's inference FLOPs at 1M tokens — which directly attacks the HBM-scarcity constraint W26 flagged. Long-context agentic workloads are becoming an open-weight strength, not a closed-model premium, so re-benchmark long-context routing assumptions.

SOURCE: DEEPSEEK V4 TECHNICAL REPORT (ARXIV 2606.19348)

PATTERN 04

Training and serving substrates diversify away from NVIDIA

`LongCat-2.0` `GPT-5.6 Sol on Cerebras` `Groq inference cloud`

In one week, a claimed frontier-scale non-NVIDIA training run (LongCat-2.0 on 50K+ domestic Chinese ASICs) and a flagship non-NVIDIA serving deal (GPT-5.6 Sol on Cerebras at up to 750 tok/s, planned July) both landed. Treat the LongCat training claim as vendor-asserted, but the direction is consistent: buyers should start asking where a model was trained and where it can be served as part of sovereignty and capacity planning.

SOURCE: SILICONANGLE; OPENAI

BENCHMARK MOVES

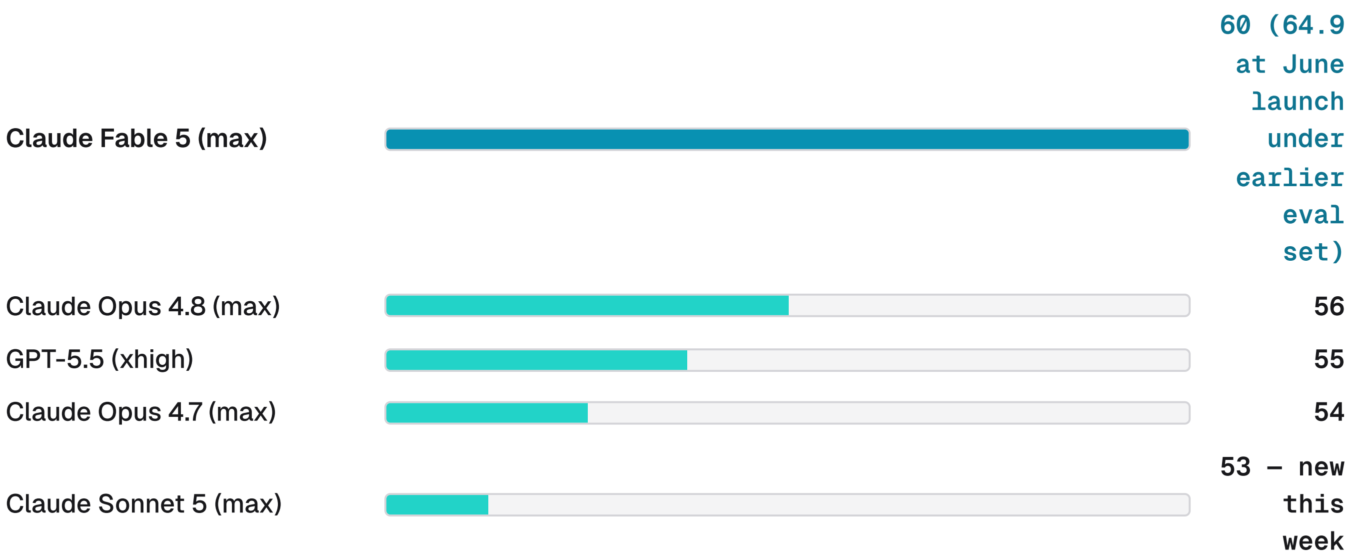
Where the leaderboard moved.

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

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX (V4.1)

MOVE 01

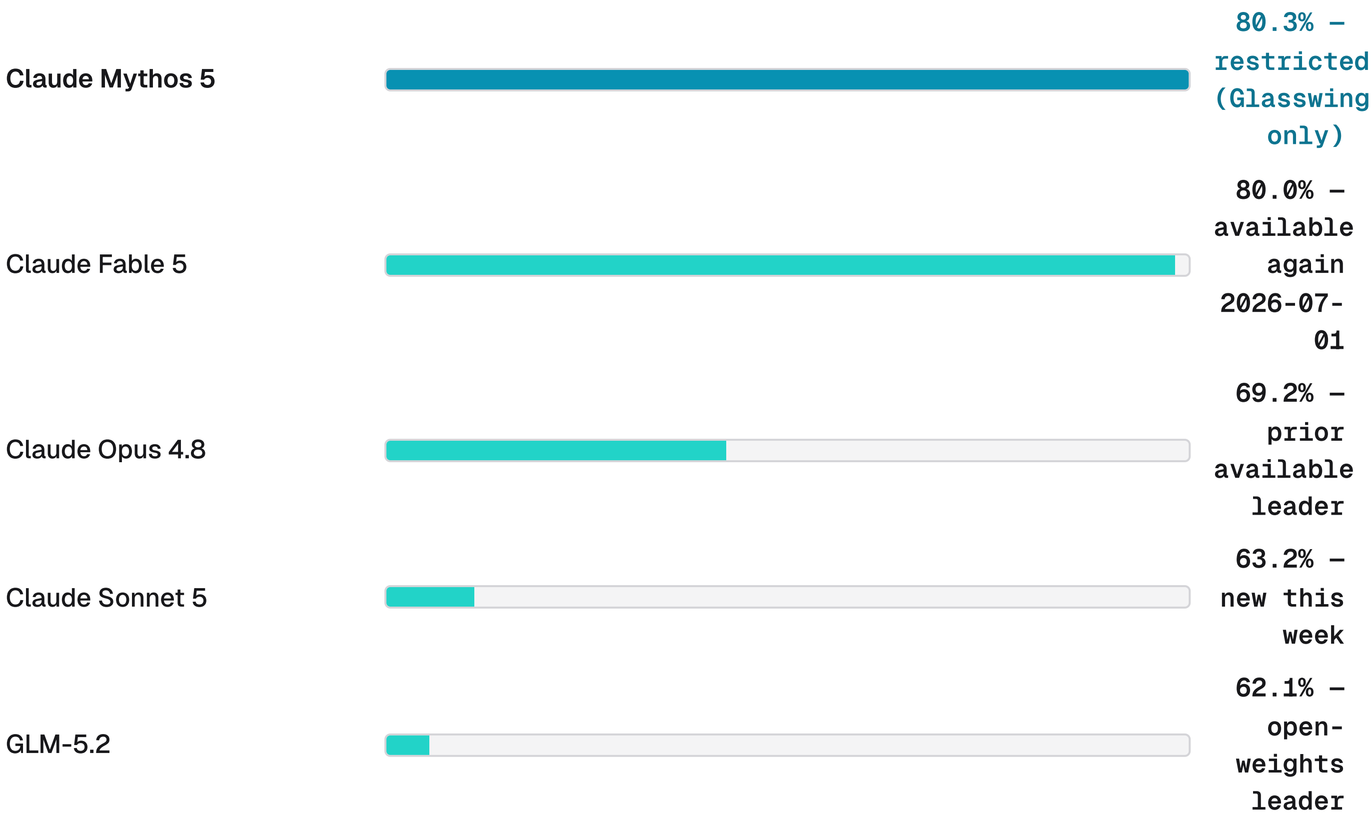
Sonnet 5 debuted at 53 (max effort), +6 over Sonnet 4.6 and tying GPT-5.5 (high); Fable 5's July 1 reinstatement puts the index leader back on sale



SOURCE: ARTIFICIAL ANALYSIS

SWE-BENCH PRO (VENDOR-SCAFFOLD AGGREGATE) MOVE 02

Sonnet 5 entered at 63.2% (+5.1pt over Sonnet 4.6, with 85.2% on Verified); Fable 5's 80.0% score re-entered the available-model conversation over the prior ~69% ceiling

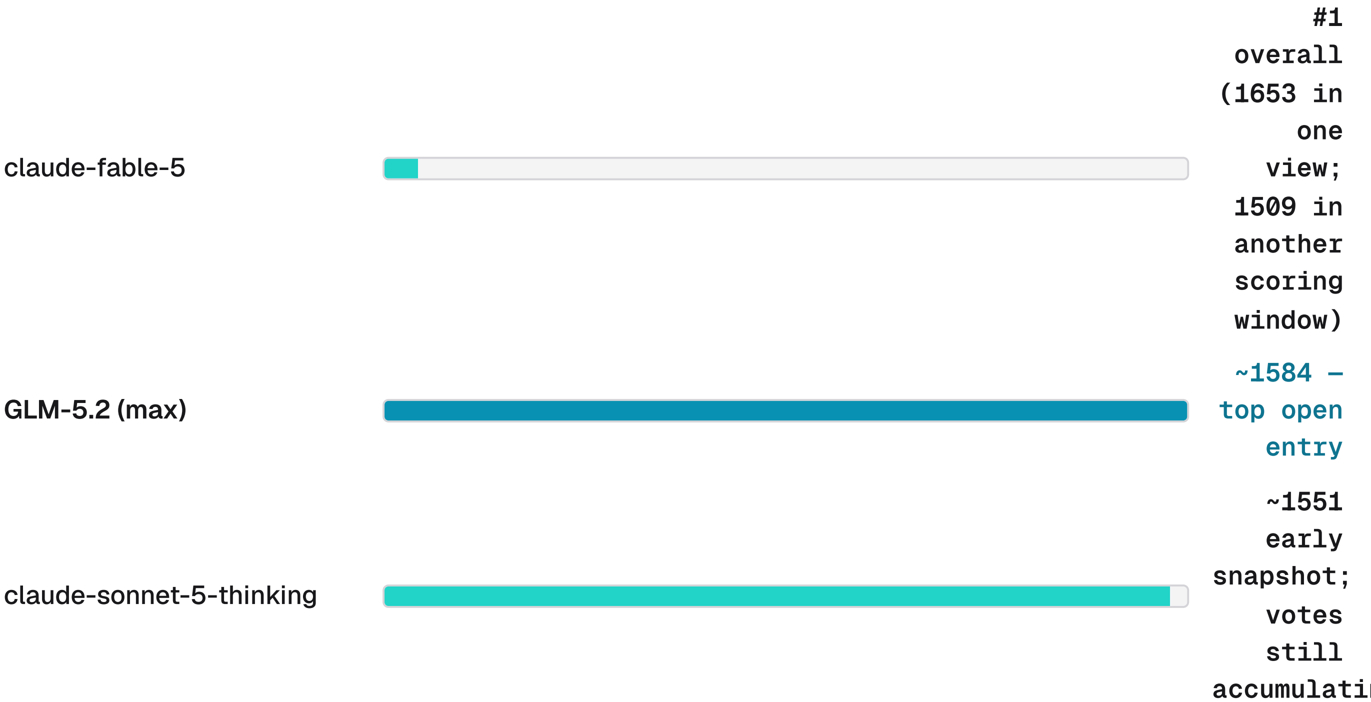


SOURCE: ANTHROPIC SYSTEM CARD VIA VELLUM; BENCHLM.AI (UPDATED 2026-07-03); VENDOR-SCAFFOLD NUMBERS, NOT SCALE-STANDARDIZED

LMARENA

MOVE 03

claude-sonnet-5-thinking added to the Code, Text, Search, Vision, and Document leaderboards on July 2; claude-fable-5 holds #1 overall while Elo settles

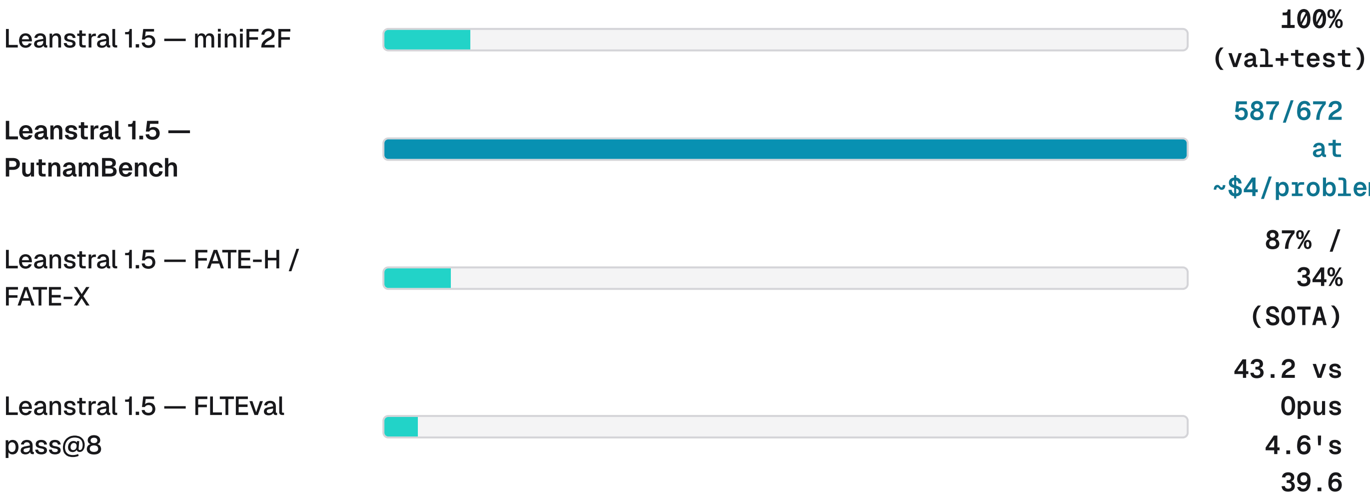


SOURCE: ARENA LEADERBOARD CHANGELOG

PUTNAMBENCH / MINIF2F (FORMAL MATH)

MOVE 04

Leanstral 1.5 reset the open state of the art: miniF2F saturated at 100%, PutnamBench 587/672 — edging Seed-Prover 1.5 high by 7 problems at ~1/75th the per-problem cost



SOURCE: MISTRAL AI (VENDOR-REPORTED; INDEPENDENT RE-RUNS PENDING)

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Jul 4, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	Claude Fable 5	Claude Opus 4.8	Fable 5's July 1 restoration makes the board leader purchasable again — but it is metered to usage credits after July 7 and its 19-day outage proved gating risk, so keep Opus 4.8 as the tested fallback.
Open frontier	GLM-5.2	LongCat-2.0	GLM-5.2 keeps the permissive open lead on scores; LongCat-2.0 arrives with real OpenRouter demand evidence but staged weights — verify availability before piloting self-host.
Reasoning	Claude Fable 5	GPT-5.5	The AA Intelligence Index leader is available again; GPT-5.5 remains the strongest generally-available OpenAI entry while GPT-5.6 sits in gated preview.
Coding	Claude Fable 5	Claude Sonnet 5	Fable 5's 80.0% SWE-bench Pro is back on sale; Sonnet 5 (63.2% Pro, 85.2% Verified) is the cost-tier challenger — priced per task, not per token.
Multimodal	Gemini 3.5 Flash	MiniMax-M3	No W27 multimodal reset; Google's media line churned (Veo 2.0/3.0 shut down June 30 as cheap successors shipped), which is a migration-budget signal rather than a leadership change.
Edge / small	Mellum2	North Mini Code	No edge leadership change in-window; TwoTower's diffusion retrofit (2.42x throughput on a 30B-class backbone) is the efficiency pattern to watch for this tier.

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.

2026-06-26
SIGNAL 01
OpenAI + Anthropic (US Commerce)

Within five days the US government gated a new frontier launch (GPT-5.6, ~20 approved partners) and un-gated a suspended one (Fable 5 restored July 1)

Frontier model availability is now partially a regulatory variable. Contracts and architectures should assume any flagship can be gated at launch or suspended post-launch; multi-model routing is availability insurance, not just cost optimization.

SOURCE: OPENAI; ANTHROPIC

2026-06-30
SIGNAL 02
Anthropic

Sonnet 5 launched at introductory $2/$10 per M tokens through August 31 (then $3/$15); Fable 5 moves to usage credits even on paid plans after July 7

Budget owners should model post-promo Sonnet 5 economics now — the independent cost-per-task data already uses standard pricing — and treat Fable 5 as a premium metered resource, not a plan entitlement.

SOURCE: ANTHROPIC

2026-07-24
SIGNAL 03
DeepSeek

Legacy deepseek-chat / deepseek-reasoner aliases go fully dark July 24 15:59 UTC; they currently silently route to deepseek-v4-flash

Anyone with DeepSeek integrations has a three-week migration deadline, and reasoner traffic wanting Pro-tier quality must explicitly move to deepseek-v4-pro. Open-weight vendors deprecate as aggressively as closed ones — budget for it.

SOURCE: DEEPSEEK API DOCS

2026-06-30
SIGNAL 04
Google

Veo 2.0/3.0 video models shut down June 30; same day, Nano Banana 2 Lite hit GA and Gemini Omni Flash entered public preview at $0.10/second of 720p video

Google's media-model line is churning fast — aggressive price points on the way in, one-year-class deprecations on the way out. Media pipelines built on Gemini need version-migration budgets as a standing line item.

SOURCE: GOOGLE GEMINI API CHANGELOG

2026-07-01
SIGNAL 05
xAI

Voice Agent Builder beta: no-code speech-to-speech agents on Grok Voice at $0.05/min audio + $0.01/min telephony — roughly 1/3 to 1/5 of ElevenLabs/Vapi-class pricing

Model vendors keep moving up-stack into the agent-platform layer, putting voice-AI middleware margins under direct attack from a model owner. Re-price any voice-agent build-vs-buy decision, and treat vendor-run voice benchmark claims as marketing.

SOURCE: XAI

2026-06-30
SIGNAL 06
Anthropic

Claude Science beta shipped — a multi-agent scientific workbench on existing Claude models, with up to 50 funded projects at $30K credits each (applications through July 15)

Anthropic is monetizing workflow ownership rather than raw capability, consistent with the W26 read that model adoption is turning into workflow operations. Expect vertical workbenches, not just model upgrades, to be the competitive surface in H2 2026.

SOURCE: ANTHROPIC; TECHCRUNCH

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.

JULY 2026
WATCH 01

GPT-5.6 general availability

Terra's claimed GPT-5.5-level capability at half the price is the biggest potential closed-model price reset of Q3; also watch whether the government preview process becomes a repeatable framework for every frontier launch.

JULY 2026
WATCH 02

Gemini 3.5 Pro GA

A 2M-context flagship GA (slipped from June) would contest the closed-frontier ordering for the first time since May; it is already on LMArena and Antigravity for some users.

JULY 7, 2026
WATCH 03

Fable 5 usage-credit transition

The included-allowance window ends; watch adoption and routing behavior once Fable 5 is purely metered, plus completeness of Bedrock/Vertex/Foundry re-enablement.

JULY 24, 2026
WATCH 04

DeepSeek legacy alias shutdown

deepseek-chat / deepseek-reasoner go dark at 15:59 UTC; expect a burst of migration issues and possible V4 usage-share data.

DAYS TO WEEKS
WATCH 05

LongCat-2.0 full weight availability and independent benchmarks

The HF repo said weights were coming soon at announcement (INT8/FP8 uploads observed since); independent evals will test both the near-frontier claim and the domestic-ASIC training story.

BY AUGUST 31, 2026
WATCH 06

Sonnet 5 promo pricing expiry

Standard $3/$15 pricing plus its measured token appetite could make Sonnet 5 more expensive per task than Opus 4.8 in production; re-run cost-per-task math before the promo lapses.

JULY 2026
WATCH 07

GPT-5.6 Sol on Cerebras

Up to 750 tok/s serving for select customers would be the first datapoint for frontier closed-model serving on non-NVIDIA silicon at scale.

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 11

Week 27 of 2026 · July 4, 2026

WEB

brianletort.ai/industry/model

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