THE MODEL PULSE ISSUE 17 · WEEK 33 OF 2026

THE BIG READ

Six releases in five days, and the model you can benchmark stopped being the model you can run

KEY SIGNALS THIS PERIOD

5	6	4	5
Models added to the tree	Vendors shipped frontier-class	Architectural patterns crossed multiple vendors	Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

August 15, 2026

Issue 17 · 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.

This was the heaviest release week of the year, and the interesting thing about it is not the capability. Meta shipped Muse Glimmer under Apache 2.0 on Monday, OpenAI shipped GPT-5.6-Cyber behind a new access tier the same day, SpaceXAI shipped Grok 4.6 on Wednesday, DeepSeek swapped its flagship endpoint to the 0813 build on Wednesday without saying so, Google shipped Gemini 3.7 Flash on Thursday at half the prior price, and Z.ai announced GLM-5.3 on Friday. Six models. Four vendors claiming some version of openness. And in three of the six cases, the artifact a buyer can actually obtain is not the artifact anyone measured.

DeepSeek is the cleanest example. On August 12 the version string on its pricing page changed from the April preview to DeepSeek-V4-Pro-0813, and the vendor-reported gains are enormous: Terminal Bench 2.1 from 72.1 to 87.9, DeepSWE from 12.8 to 62.7, CyberGym from 52.7 to 83.3, the Artificial Analysis Intelligence Index from 45 to 53. There was no blog post. The Hugging Face repositories continued to host the April preview. For a stretch this week, every independent evaluation of 'DeepSeek V4 Pro' was measuring a build nobody could download, and every downloadable build was one nobody was serving. Z.ai did a softer version of the same thing: GLM-5.3 was announced with a claimed best-in-open-source Terminal Bench 3.0 result and a CyberGym score ahead of Claude Mythos 5, reachable at announcement only through a paid coding subscription, with weights promised on Hugging Face 'within two weeks.'

This publication has recorded both models as gated rather than open-weight in the tree this week. That is not pedantry. Last week's issue documented Artificial Analysis measuring identical open weights losing up to 40% of reference accuracy depending on which endpoint served them. Stack the two findings and the procurement problem is concrete: the score belongs to a specific build served in a specific configuration, and a buyer downloading weights under the same product name is acquiring neither. The gap used to be closed-versus-open. The gap that matters now is served-versus-downloadable, and it opens and closes on a vendor's release schedule rather than on a license.

Meta went the other direction and it is worth being precise about what it did. Muse Glimmer is a 30B dense multimodal model under Apache 2.0 — no 700-million-user gate, no naming rule, no acceptable use policy, and no mechanism for Meta to withdraw the grant. Artificial Analysis puts it at 35 on the Intelligence Index and 44 on the Openness Index. But Meta published weights and withheld training data and training code, and it is not serving the model on its own API, so every price and latency number a buyer sees for Glimmer comes from a third party. Meta gave away the terms and kept the frontier: Muse Spark 1.2 stays closed, with its weights promised later.

The pricing story runs the same shape. Headline rates fell — Gemini 3.7 Flash at $0.75 / $3.75 is half of 3.6 Flash, and Grok 4.6 ties GPT-5.6 Sol Max on the Intelligence Index at $2 / $6 against Sol's $5 / $30. But Google's own pricing page says the Gemini rate is introductory through December 31 and that $1.50 / $7.50 applies from January 1, and DeepSeek's flat $0.435 / $0.87 gives way on August 16 to reported peak and off-peak tiers of $1.32 / $3.96 and half that off-peak. Both increases were disclosed in advance, in writing, by the vendor. Anyone modelling agent unit economics on this week's rate card is modelling a promotional period with a published expiry date, and the expiries are inside the planning horizon for anything being procured this quarter.

The one number that survives all of this is Grok 4.6's turn count. Artificial Analysis measured roughly 53 turns and 0.5B input tokens to resolve long-horizon agentic tasks against roughly 103 turns and 2.0B tokens for Claude Opus 5 at maximum settings. Turn efficiency is a property of the model that does not reprice on August 16 and does not depend on which endpoint serves it. For anyone running agents at scale, that is the more durable procurement input than the rate card — and it is the number the fewest buyers are tracking.

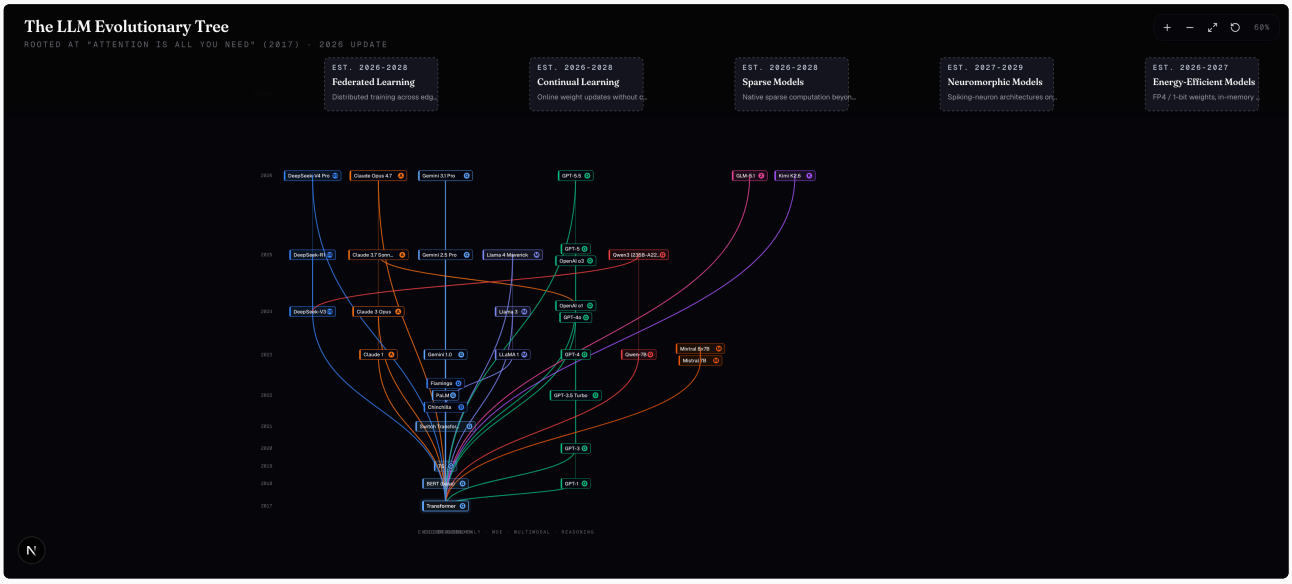
THE BAR

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.

5 model rows added and 1 updated — the heaviest single week the tree has recorded in 2026. Two of the five additions are classified gated rather than open_weights despite open-source framing in their own announcements, and the updated row publishes weights for a build its API no longer serves.



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

ADDED (5)

- muse-glimmer-30b
- gpt-5-6-cyber
- grok-4-6
- gemini-3-7-flash
- glm-5-3

UPDATED (1)

- deepseek-v4-pro

The openness classification did more work this week than the branch classification. GLM-5.3 and GPT-5.6-Cyber are both recorded as gated: one because its weights are promised but unshipped, the other because access runs through a vetting tier with hardware security keys. Reclassify GLM-5.3 to open_weights when the Hugging Face release is confirmed — the tree records what shipped, not what was announced.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-08-12

RELEASE 01

SpaceXAI

FRONTIER

REASONING

Grok 4.6

Ties GPT-5.6 Sol Max at 61 on the Intelligence Index at $2 / $6, and resolves long-horizon agentic tasks in roughly half the turns of Claude Opus 5

Explicitly a post-training release on the Grok 4.5 foundation rather than a bigger base model, five weeks after its predecessor. The tie on the Index is the least useful number in the release: architects should look at the measured ~53 turns and ~0.5B input tokens against Claude Opus 5's ~103 turns and ~2.0B at max settings, because turn count drives the bill on any long-running agent and does not move when a rate card changes. The gap it did not close is Terminal-Bench v3.0, where 26% against GPT-5.6 Sol Max's 34.6% still argues against putting it on unattended terminal work.

SOURCE: VENTUREBEAT, ARTIFICIAL ANALYSIS, THE AGENT REPORT

2026-08-13

RELEASE 02

Google

FRONTIER

MULTIMODAL

Gemini 3.7 Flash

1M context workhorse at $0.75 / $3.75, half of 3.6 Flash — with a doubling to $1.50 / $7.50 published at launch for January 1

Three weeks after 3.6 Flash, which is a cadence that tells you where Google is putting its attention: the Pro line has no announced timeline and Pichai declined Pro-cadence questions on the last earnings call. Buyers should treat the price as the story and read the footnote — Google's own pricing page states the rate is introductory through December 31. This is the clearest public admission by any lab that current agent unit economics are promotional, and it is disclosed rather than discovered, which is the honest way to do it.

SOURCE: GOOGLE BLOG, GOOGLE CLOUD PRICING PAGE, INFOWORLD

2026-08-10

RELEASE 03

OpenAI

SPECIALIST

REASONING

GPT-5.6-Cyber

Purpose-trained security model gated behind a new Daybreak Red tier at $12.50 / $75, with hardware security keys required from September 1

Reported to answer 95.0% of advanced cyber requests against 1.5% for GPT-5.6 Sol under normal safeguards, and rated High rather than Critical under OpenAI's Preparedness Framework. The release is a distribution decision more than a capability one: Daybreak Blue gives approved defenders general-purpose models with cyber guardrails removed, Red gates the purpose-trained models behind vetting. Security leaders should read this as the first frontier model where the access-control design is the product, and plan for the vetting timeline rather than the API key. Recorded at medium placement confidence: no first-party OpenAI source was located during research.

SOURCE: AI/TLDR, DEV COMMUNITY, THE NEURON

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.

2026-08-10

RELEASE 01

Meta

EDGE / SMALL

DENSE

Muse Glimmer

30B dense multimodal under Apache 2.0 — Meta's first open release with no user gate, no naming rule and no acceptable use policy

Artificial Analysis scores it 35 on the Intelligence Index and 44 on the Openness Index, level with DeepSeek V4 Flash and GLM-5.2. The capability is mid-tier for its class; the license is the release. Every prior Meta open model carried a 700-million-user gate and a 'Built with Llama' naming requirement, and Apache 2.0 leaves Meta no lever to withdraw the grant — so teams that rejected Llama on legal review should re-run that review. Two caveats worth carrying into the decision: Meta withheld training data and code, and it serves no API for the model, so every price and latency figure comes from a third party.

SOURCE: META AI RESEARCH, ARTIFICIAL ANALYSIS, VENTUREBEAT

2026-08-12

RELEASE 02

DeepSeek AI

OPEN FRONTIER

MOE

DeepSeek-V4 Pro (0813)

Flagship leaves preview with large vendor-reported gains — announced by nothing but a version string, and with the April weights still the only ones downloadable

Terminal Bench 2.1 from 72.1 to 87.9, DeepSWE from 12.8 to 62.7, CyberGym from 52.7 to 83.3, Intelligence Index from 45 to 53 — all vendor-reported, all on a build that was not on Hugging Face when the endpoint swapped. Buyers running V4 Pro through the API got a materially different model this week without a release note; buyers self-hosting got nothing. Anyone with a benchmark result for 'DeepSeek V4 Pro' in a procurement document should date it and note which build it refers to, because the name now covers two materially different artifacts.

SOURCE: UNITE.AI, DECRYPT, OPENLLMSTACK, OFLIGHT

2026-08-14

RELEASE 03

Z.ai

OPEN FRONTIER

MOE

GLM-5.3

Same 753B MoE as GLM-5.2 with a long-horizon post-training run — best open-source Terminal Bench 3.0 claimed, weights promised within two weeks

The architecture did not change at all; the delta is post-training inside sandboxes built to mimic developer workstations, with some tasks running for days. That is a useful data point on how much long-horizon capability is now recoverable from post-training alone on a fixed base. Z.ai reports a CyberGym score ahead of Claude Mythos 5 while trailing it on two other cybersecurity benchmarks, which is a narrower claim than the headline suggests. Recorded as gated: at announcement the only access was a paid coding subscription, so treat the open-source framing as an intention until the Hugging Face release lands.

SOURCE: SILICONANGLE, Z.AI

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

Post-training is the release; the base model stopped moving

Grok 4.6 | GLM-5.3 | DeepSeek-V4 Pro 0813

Three of this week's six releases explicitly ship the same foundation as their predecessor with a longer or differently-shaped post-training run. SpaceXAI said so directly for Grok 4.6 — regenerated supervised fine-tuning trajectories plus reinforcement learning in agentic environments — and Z.ai's GLM-5.3 is architecturally identical to GLM-5.2 at 753B MoE and 1M context. The procurement consequence is that version numbers no longer imply retraining cost or a capability step change, and the release cadence a vendor can sustain is now decoupled from its pretraining compute. Architects sizing a model refresh cycle should stop assuming a new point release means a new base.

SOURCE: THE AGENT REPORT, SILICONANGLE, OFLIGHT

PATTERN 02

The served build and the published weights have separated

DeepSeek-V4 Pro 0813 | GLM-5.3 | Muse Glimmer

DeepSeek swapped its production endpoint to a new build while Hugging Face continued to host the prior one. Z.ai benchmarked a model available only by subscription and promised weights later. Meta published weights for a model it does not serve at all, so the artifact is downloadable but has no reference endpoint. In each case the name on the benchmark and the name on the download resolve to different things, which breaks the assumption underneath every open-weight evaluation: that a published score describes an artifact you can obtain. Combined with last week's Endpoint Accuracy Index finding of up to 40% accuracy loss by serving configuration, a model score is now a claim about a build-and-endpoint pair, not about a model.

SOURCE: UNITE.AI, SILICONANGLE, ARTIFICIAL ANALYSIS

PATTERN 03

Published price expiries replace silent repricing

Gemini 3.7 Flash | DeepSeek-V4 Pro | GPT-5.6-Cyber

Google states on its pricing page that Gemini 3.7 Flash's $0.75 / $3.75 runs through December 31 and becomes $1.50 / $7.50 on January 1. DeepSeek warned of a significant increase and then landed peak and off-peak tiers effective August 16, reported at $1.32 / $3.96 peak against a flat $0.435 / $0.87 before. Both are disclosed in advance rather than discovered in an invoice, which is a genuine improvement in vendor behavior — and also the clearest signal yet that the current cost of inference is subsidized. Finance teams should build the published step-up into any agent business case whose payback period crosses January.

SOURCE: GOOGLE CLOUD PRICING, OPENLLMSTACK, INFOWORLD

PATTERN 04

Dense at 30B as the local-agent target, against the MoE consensus

Muse Glimmer | LFM2.5-2.6B | Qwen3.6-27B

Nearly every frontier release this year has been a mixture-of-experts design, and Muse Glimmer is deliberately not: 30B parameters that all activate on every forward pass, shipped with two 4-bit quantizations and a speculative-decoding drafter at a roughly 20 GB deployment envelope. Dense is the easier target for consumer runtimes and quantization toolchains, which is the point when the deployment surface is a single 24-32 GB consumer GPU rather than a datacenter. The pattern to watch is a bifurcation: MoE where memory bandwidth is cheap and dense where the constraint is a laptop.

SOURCE: META AI RESEARCH, MINDSTUDIO, ARTIFICIAL ANALYSIS

BENCHMARK MOVES

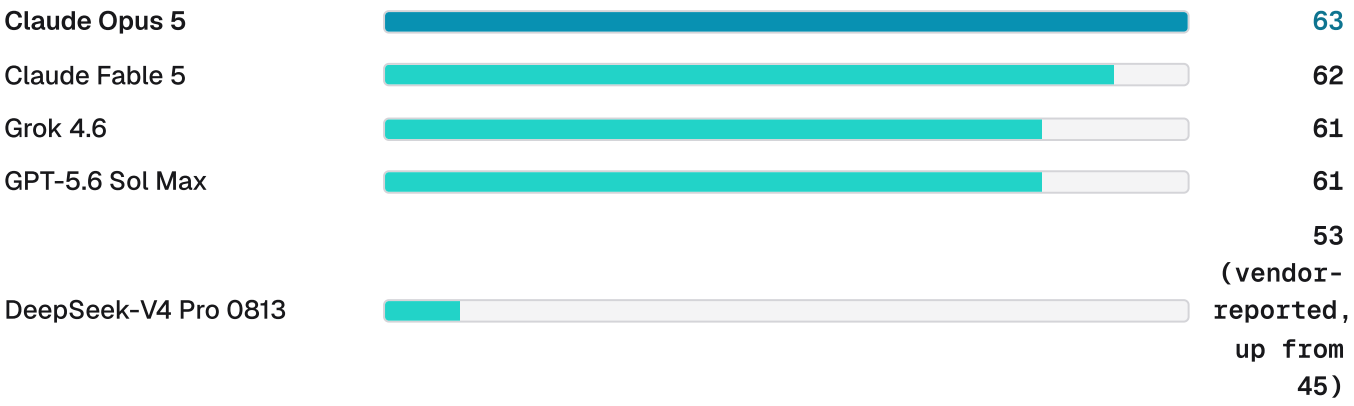
Where the leaderboard moved.

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

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX

MOVE 01

Grok 4.6 enters at 61 on a post-training release alone, tying GPT-5.6 Sol Max and putting four models within two points of the lead



SOURCE: ARTIFICIAL ANALYSIS VIA VENTUREBEAT AND THE AGENT REPORT

GDPVAL-AA V2 (REAL-WORLD KNOWLEDGE WORK, ELO)

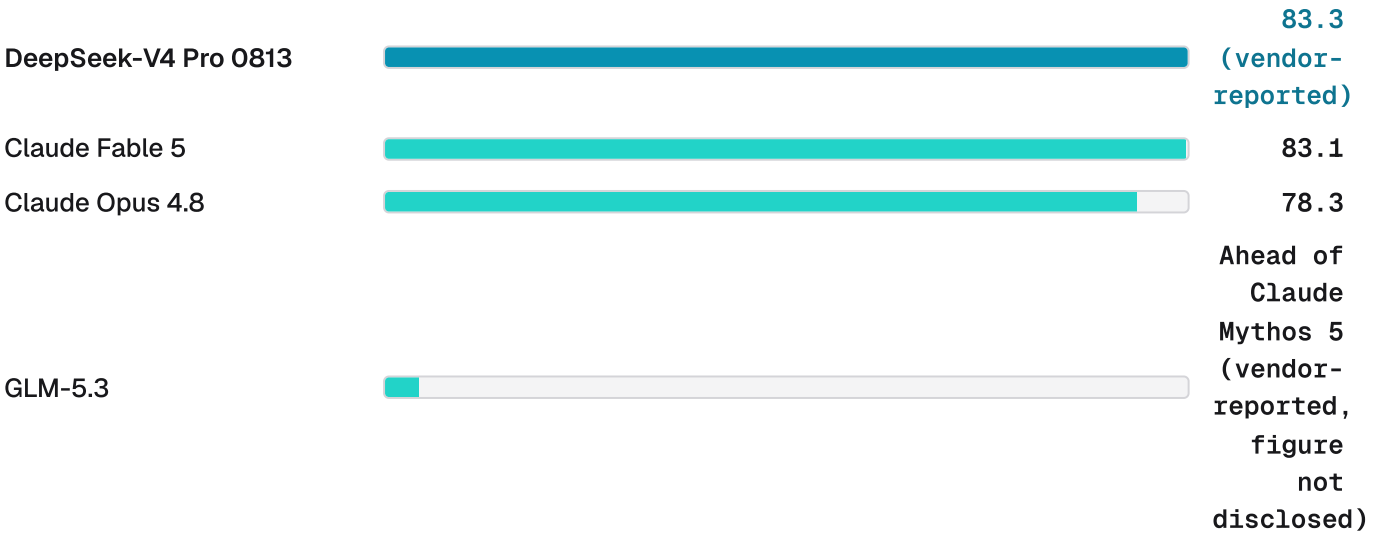
MOVE 02

Grok 4.6 takes the top position at 1,753, a 227-point jump over Grok 4.5 and the largest single-release move on this board this year

SOURCE: ARTIFICIAL ANALYSIS VIA VENTUREBEAT

CYBERGYM (VULNERABILITY DISCOVERY) MOVE 03

Two separate vendors claimed the top spot on the same board in the same week, one open-weight and one Chinese subscription-gated — and neither claim has been independently reproduced



SOURCE: DEEPSEEK VIA OFLIGHT AND DECRYPT; Z.AI VIA SILICONANGLE

TERMINAL-BENCH (V2.1 AND V3.0 — NOT COMPARABLE TO EACH OTHER) MOVE 04

The board split into two versions being quoted interchangeably, which is where most of this week's coverage went wrong

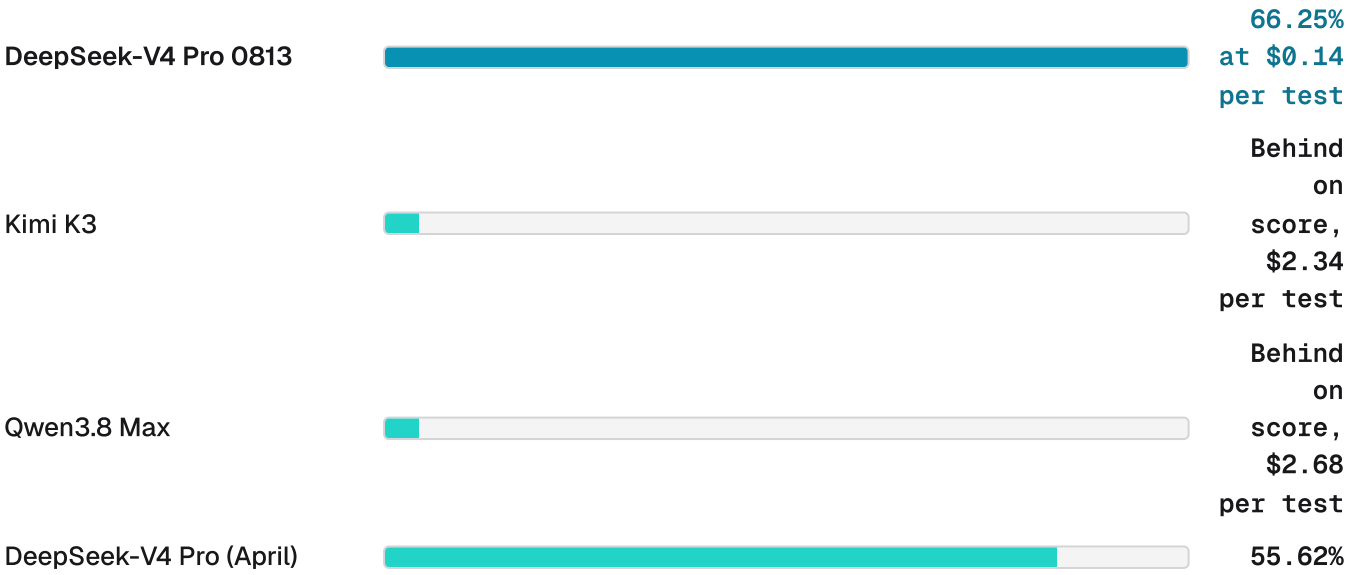
DeepSeek-V4 Pro 0813 (v2.1)	87.9 (vendor-reported, up from 72.1)
Claude Fable 5 (v2.1)	88.0
GPT-5.6 Sol Max (v3.0)	34.6%
Grok 4.6 (v3.0)	26%
GLM-5.3 (v3.0)	Highest open-source score claimed (vendor-reported)

SOURCE: OFLIGHT, VENTUREBEAT, SILICONANGLE

VALS INDEX, OPEN-WEIGHTS VIEW

MOVE 05

DeepSeek-V4 Pro 0813 moves to second at 66.25% from the April build's 55.62%, and the cost column separates it further than the score does



SOURCE: VALS AI VIA OPENLLMSTACK

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Aug 15, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Opus 5**	Grok 4.6	Opus 5 holds the Intelligence Index lead at 63, but Grok 4.6 tied GPT-5.6 Sol Max at 61 this week on a post-training run alone while charging $2 / $6 against Sol's $5 / $30. Four models now sit within two points, so the tier is decided on cost per completed task rather than on the Index.
Open frontier	**DeepSeek-V4 Pro 0813**	Kimi K3	Leader on served scores and unbeatable on price at $0.14 per Vals test, with an asterisk this publication is applying for the first time: the 0813 weights were not downloadable when the endpoint swapped, so the leading open model was briefly not obtainable. GLM-5.3 may take this row within two weeks if its promised weights land.
Reasoning	**Claude Opus 5**	GPT-5.6 Sol Max	Unchanged on capability, but the efficiency read moved against Opus 5 this week: Artificial Analysis measured it taking roughly 103 turns and 2.0B input tokens on long-horizon agentic tasks against Grok 4.6's ~53 turns and ~0.5B for a comparable answer quality.
Coding	**Claude Fable 5 Max**	GPT-5.6 Sol Max	Fable 5 Max leads CursorBench v3.2 at 70.5% and FrontierCode v1.1 Extended at 63.6%, while GPT-5.6 Sol Max takes DeepSWE v1.1 at 73%. Grok 4.6 closed most of the gap on DeepSWE this week, rising from 54% to 65.9%, without taking either row.
Multimodal	**Gemini 3.7 Flash**	Muse Spark 1.2	Google took this row on economics rather than capability: 1M context with text, image, audio and video input at $0.75 / $3.75 is half the prior Flash rate, and the model went straight into Gemini Spark. Read the January 1 doubling to $1.50 / $7.50 before treating it as the durable price.
Edge / small	**Muse Glimmer**	Qwen3.6-27B	Muse Glimmer takes the row on licensing, not on score — 35 on the Intelligence Index is mid-pack for its

TIER	LEADER	CHALLENGER	READ
			class, but Apache 2.0 with no user gate, naming rule or acceptable use policy removes the legal friction that kept Llama out of many production pipelines. Liquid's LFM2.5-2.6B remains the pick well below 20 GB.

VENDOR SIGNALS

Pricing, gating, deprecation.

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

2026-08-12

SIGNAL 01

DeepSeek AI

Flagship endpoint swapped to the 0813 build with no announcement, and a price increase scheduled for August 16 replacing the flat rate with peak and off-peak tiers

Two procurement problems in one move. The silent build swap means any team with a pinned expectation of model behavior got a different model without notice, so anyone running DeepSeek in production should re-run their own evals this week. The pricing change, reported at $1.32 / $3.96 peak against $0.435 / $0.87 before, lands the day after this issue publishes and materially changes the cost case that made V4 Pro the default cheap frontier option.

SOURCE: UNITE.AI, DECRYPT, OPENLLMSTACK, OFLIGHT

2026-08-10

SIGNAL 02

Meta

First open release under Apache 2.0, abandoning the Llama License's user gate, naming rule and acceptable use policy, with Muse Spark 1.2 weights promised to follow

Legal and procurement teams that blocked Llama-family models on license terms have lost their objection for this model, and should re-run the review rather than carry forward a stale decision. The strategic read is that Meta stopped charging in terms — a price it could only charge while buyers had no substitute — and the promised Muse Spark 1.2 weights would extend that to a frontier-class model.

SOURCE: META AI RESEARCH, VENTUREBEAT, BUSINESS MODEL ANALYST

2026-08-10

SIGNAL 03

OpenAI

Daybreak program split into Blue and Red tiers, with hardware security keys required for login from September 1

The first frontier lab to make physical authentication a condition of model access. Security teams planning to use Daybreak-tier models for vulnerability research should start the vetting and key-provisioning process now rather than at the point of need, because the gate is organizational rather than technical and the September 1 date is firm.

SOURCE: AI/TLDR, DEV COMMUNITY

2026-08-13

SIGNAL 04

Google

Third Flash release in roughly six weeks while the Pro line has no announced timeline and Pichai declined Pro-cadence questions on the last earnings call

Google is iterating fast where the volume is and slowly where the headlines are. Architects should not wait on a Pro refresh to make agent platform decisions this quarter, and should note that the Flash tier now carries Pro-level agentic positioning in Google's own documentation.

SOURCE: GOOGLE BLOG, INFOWORLD

2026-08-14

SIGNAL 05

Z.ai

GLM-5.3 benchmarked and announced as open-source with weights promised on Hugging Face within two weeks, available at launch only through a paid coding subscription

A benchmark-first, weights-later release pattern that is becoming common enough to plan around. Teams should not schedule a self-hosted deployment against an announcement date, and should treat the two-week window as a forecast rather than a commitment — the tree will reclassify the model only when the weights are confirmed.

SOURCE: SILICONANGLE

WATCHLIST

On the radar next.

6 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.

AUG 16

WATCH 01

DeepSeek peak / off-peak pricing takes effect

The flat $0.435 / $0.87 rate ends and reported peak tiers of $1.32 / $3.96 begin, with peak hours at 01:00-04:00 and 06:00-10:00 UTC. Watch whether workloads visibly shift into off-peak windows, which would be the first evidence of time-of-day arbitrage becoming a real inference architecture pattern.

AUG 16-28

WATCH 02

GLM-5.3 weights on Hugging Face

Z.ai promised weights within two weeks. If they land, the model likely takes the open frontier row and this publication reclassifies it from gated. If they slip, the benchmark-first release pattern is worth treating as a durable vendor behavior rather than a one-off.

AUG 16-31

WATCH 03

DeepSeek-V4-Pro-0813 weights, and independent reproduction of its gains

Every headline number for the 0813 build is vendor-reported. The Terminal Bench 2.1 jump from 72.1 to 87.9 and DeepSWE from 12.8 to 62.7 are large enough that independent replication is the only thing that should move them into a procurement document.

SEP 1

WATCH 04

OpenAI hardware security key requirement for Daybreak login

The first hard physical-authentication gate on frontier model access. Whether other labs follow within the quarter determines if this becomes a category norm for dual-use capability or stays an OpenAI-specific control.

AUG 16-SEP 15

WATCH 05

Muse Spark 1.2 open weights

Zuckerberg said the weights are coming. A frontier-class Meta model under a permissive license would be the largest single addition to the open tier this year and would force a re-read of the open-versus-closed capability gap lever.

SEP-OCT

WATCH 06

Independent turn-efficiency measurement across the frontier

Grok 4.6's ~53-turn result against Claude Opus 5's ~103 is currently a single-source measurement on a small set of tasks. If Artificial Analysis or another independent lab extends turn counting across the full frontier, it becomes the most decision-relevant number in agent procurement.

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 17

Week 33 of 2026 · August 15, 2026

WEB

brianletort.ai/industry/model

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