THE MODEL PULSE ISSUE 16 · WEEK 32 OF 2026

THE BIG READ

Two new frontier models shipped and the more important releases were the measuring instruments — the same weights now score differently depending on who serves them

KEY SIGNALS THIS PERIOD

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

AUTHOR
Brian Letort
BrianLetort.AI

PUBLISHED
August 8, 2026
Issue 16 · 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.

Artificial Analysis launched the Endpoint Accuracy Index on Aug 4 and it reframes every open-weights procurement decision this publication has made. The method is explicit: benchmark each serverless endpoint against a self-hosted reference deployment of the official weights at the lab's recommended precision, across tool calling (BFCL-500), scientific reasoning (HLE-250) and long-context recall (AA-LCR-25). The findings are not marginal. Some gpt-oss-120b endpoints score 22% on BFCL-500 against 37% for the reference, driven by tool-call parsing and formatting differences. The most restrictive GLM-5.2 endpoints score half the reference or less on HLE-250 because output-token limits truncate reasoning before it completes. DeepSeek V4 Pro is the good case — most endpoints at parity, with DeepSeek's own first-party endpoint slightly above reference. Endpoints scoring below reference generally produce fewer output tokens per task, and the worst produce about half. A model ID is no longer sufficient to specify what you bought.

Two days later the same firm published Intelligence Index v4.1.1, a methodology patch that moved HLE, AA-LCR and AA-Omniscience grading to GPT-5.6 Luna, unifying three evaluations under one modern grader. Most models moved under a point. The largest mover was Muse Spark 1.2 at +2.7 — almost as much as the model gained from its own release (51 to 54 from Muse Spark 1.1). Pin the Index version in any comparison you carry forward, because a grader change is now capable of moving a model nearly as much as a generation.

The actual releases were Meta's Muse Spark 1.2 (Aug 5) and Alibaba's Qwen3.8-Max API GA (Aug 3), and both tell a cost story rather than a capability story. Muse Spark 1.2 rose to Index 54 pre-patch and posted a GDPval-AA v2 Elo of 1631, up 260 points from 1.1, at unchanged list pricing of $1.25/$4.25 — yet Artificial Analysis measured cost per Index task rising from $0.29 to $0.40, roughly 38%, because input tokens rose ~53% and output tokens ~36% per task. Qwen3.8-Max landed at Index 58 and consumed 150M output tokens to run the Index against a

class median of 66M, which AA labels 'very verbose'. Rate cards did not move; bills did.

Two honesty notes. Muse Spark 1.2's AA-Omniscience gain (18 to 22) comes with hallucination falling from 38% to 28% while the attempt rate fell from 82% to 67% and accuracy fell from 41% to 38% — the model is refusing more, not knowing more, and the index does not penalize refusal. And Liquid's LFM2.5-2.6B shipped on Aug 4 with a release page describing deployment 'without restrictions' while the LICENSE in the same repository withholds commercial use from any entity above $10M in revenue. Net/net: pin the endpoint, pin the Index version, read the license, and measure cost per completed task rather than per token.

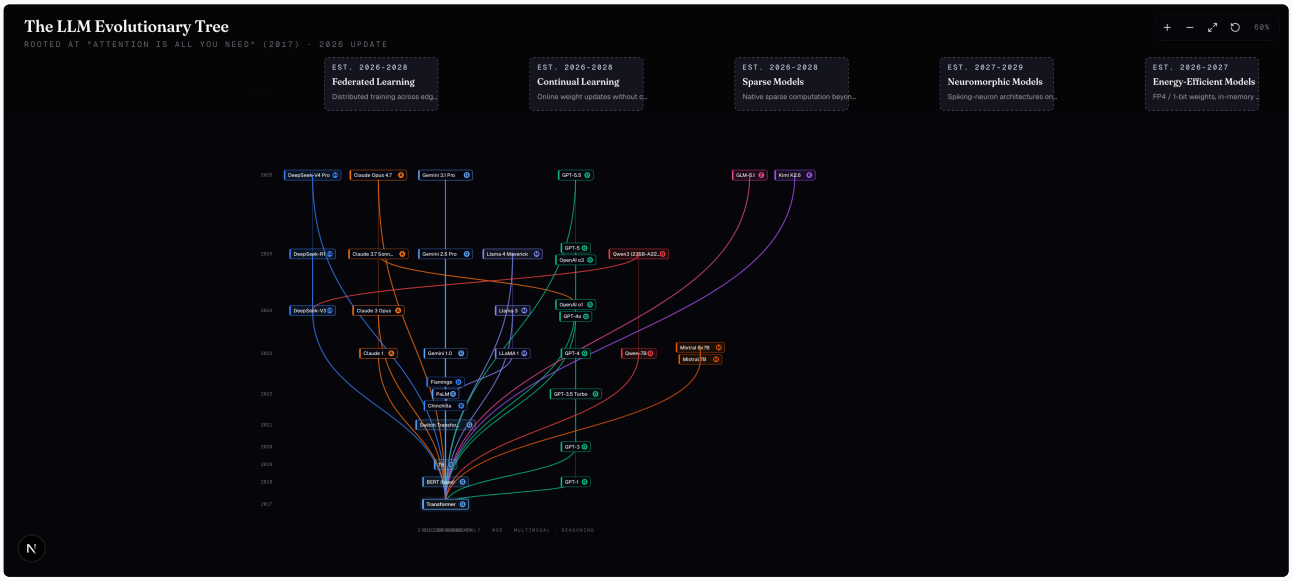
THE BAR

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.

Muse Spark 1.2 and Qwen3.8-Max join as closed frontier rows, and LFM2.5-2.6B joins the edge branch as the first entry recorded with a license-contradiction annotation rather than a clean open-weight classification. Kimi K3 updates to carry Artificial Analysis's new 'Open Weights (Commercial Use Restricted)' category label, which the firm introduced this week because revenue-gated open weights have become common enough to institutionalize.



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

ADDED (3)

- muse-spark-1-2
- qwen3-8-max
- lfm2-5-2-6b

UPDATED (1)

- kimi-k3

Qwen3.8-Max enters with a disclosed conflict rather than a settled spec: Alibaba's launch blog describes 2.4T MoE with 95B active parameters while Artificial Analysis's model page states Alibaba has not disclosed model size or parameter count. The row records the vendor figures as vendor-stated and uncorroborated. Open weights for the Qwen-Max class and a companion Qwen3.8-27B were announced as coming 'next week' and had not shipped as of Aug 8, so neither is added — the W30 Kimi K3 lesson applies. OpenAI's Astra is deliberately excluded: it has not shipped, and its only public capability signal is a safety flag rather than a benchmark.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-08-05

RELEASE 01

Meta

FRONTIER REASONING

Muse Spark 1.2

Index 54 pre-patch (+3 over 1.1) and GDPval-AA v2 Elo 1631, up 260 points — at a 38% higher cost per task on unchanged list pricing

Re-baseline on cost per completed task, not the rate card: pricing held at $1.25/$4.25 per 1M with $0.15 cached input, but Artificial Analysis measured cost per Index task rising from $0.29 to $0.40 because input tokens rose ~53% and output ~36% per task, concentrated in GDPval. Terminal-Bench v2.1 moved 78% to 80% and τ³-Banking 25% to 27%, while SciCode fell 58% to 56% and HLE 45% to 44%. Meta gave AA pre-release access, and Meta's own coding comparison ran 1.1 in mini-swe-agent against 1.2 in Muse Code, so part of the vendor-reported coding delta is harness rather than model.

SOURCE: META AI RESEARCH; ARTIFICIAL ANALYSIS (INDEPENDENT); META DEVELOPER PRICING DOCS

2026-08-03

RELEASE 02

Alibaba / Qwen

FRONTIER MOE

Qwen3.8-Max

API GA at Index 58 with 1M context and $2.00/$6.00 pricing — and 150M output tokens to run the Index against a 66M class median

Treat verbosity as the procurement risk here: AA labels it 'very verbose' and spent $1,741.41 evaluating it on the Index, so blended token assumptions built from other models will understate the bill. Cache hits at $0.25 cut input cost 88%, which makes prompt-caching discipline unusually valuable on this model. Reasoning model with text, image and video input; blended 7:2:1 rate is $1.18 per 1M. Vendor claims 2.4T MoE with 95B active parameters, but AA states Alibaba has not disclosed size — treat parameters as uncorroborated.

SOURCE: ARTIFICIAL ANALYSIS MODEL PAGE (INDEPENDENT); ALIBABA CLOUD LAUNCH BLOG (VENDOR)

2026-08-07

RELEASE 03

OpenAI

REASONING AGENTIC

Astra (unreleased)

Launch delayed after internal evaluations could not rule out a Critical cybersecurity capability level — the first such flag OpenAI has made

Do not put Astra in a roadmap slot yet. OpenAI reported 'significant advancements in agentic coding and cybersecurity' strong enough that it cannot rule out Critical under its Preparedness Framework, paused internal activities not meeting stricter controls, and Altman confirmed the delay. No benchmark numbers, no ship date, and no final rating exist. The Framework's stated policy at Critical is to halt development, and OpenAI paused selectively rather than halting — which implies its own assessment is that the threshold is not yet met.

SOURCE: OPENAI; THE DECODER

2026-08-06

RELEASE 04

OpenAI

FRONTIER

REASONING

GPT-5.6 Sol / Luna ChatGPT surface update

Product-surface change only: ChatGPT chat moves to August Sol/Luna variants while Codex and ChatGPT Work stay on the July checkpoints

This is a naming hazard, not a capability event. The same model name now refers to different checkpoints depending on which surface serves it, which breaks the assumption that an internal evaluation of 'GPT-5.6 Sol' transfers across products. No API checkpoint changed and no API price changed. Safety designation is unchanged at High for Cyber and Bio/Chem. Vendor-internal factual-error reduction claims circulated via secondary press; treat as vendor-stated.

SOURCE: OPENAI DEPLOYMENT SAFETY HUB; SECONDARY PRESS FOR THE INTERNAL STATISTICS

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.

2026-08-04

RELEASE 01

Liquid AI

EDGE / SMALL

DENSE

LFM2.5-2.6B

2.6B on ~34T tokens with 128K context and strong on-device throughput — under a license withholding commercial use above $10M revenue

Send this to counsel before any pilot. The release page says 'Open-weight — Download, fine-tune, and deploy without restrictions'; Section 5(b) of the LFM Open License v1.0 in the same repository says commercial use by an entity exceeding the $10M annual revenue Threshold is 'not licensed under this Agreement', Commercial Use reaches internal enterprise use, Legal Entity includes the controlled group, and Section 11 terminates automatically on non-compliance. This is materially stricter than Kimi K3's W31 terms, which gated hosting above $20M and pointed to a separate agreement — the LFM text names no such path. On capability: base and post-trained checkpoints, four-stage post-training ending in agentic RL inside real harnesses, 220 tok/s decode on an M5 Max and roughly 30 tok/s on a phone under 2.5 GB.

SOURCE: LIQUID AI BLOG; HUGGING FACE MODEL CARD; LFM OPEN LICENSE V1.0 TEXT

2026-08-03

RELEASE 02

Alibaba / Qwen

OPEN FRONTIER

MOE

Qwen-Max class open weights and Qwen3.8-27B (announced, not shipped)

Announced as coming 'next week' alongside the Qwen3.8-Max API launch — no artifacts and no license text as of Aug 8

Do not build a self-host plan on this. Nothing shipped in-window, the license is undisclosed, and Artificial Analysis's FAQ states plainly that Qwen3.8 Max is not open source. Track it as a promise with a deadline rather than an available artifact, and note that the license terms matter more than the weights given how the last three revenue-gated releases have read.

SOURCE: ALIBABA CLOUD LAUNCH BLOG; ARTIFICIAL ANALYSIS MODEL PAGE FAQ

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 serving endpoint is now an architecture variable, not a delivery detail

`gpt-oss-120b: some endpoints at 22% on BFCL-500 against a 37% self-hosted reference`

`GLM-5.2: most restrictive endpoints at half the reference or less on HLE-250 from output-token caps`

`DeepSeek V4 Pro: majority of endpoints at parity, first-party endpoint slightly above reference`

Artificial Analysis identified two concrete mechanisms rather than asserting variance. Tool-call parsing and formatting differ by provider, which is what moves BFCL-500. Output-token limits truncate reasoning before it completes, which is what moves HLE-250 — and the firm notes that endpoints scoring below reference generally produce fewer output tokens per task, with the worst producing roughly half. For architects the consequence is that an inference provider is now part of the model's effective configuration, and any evaluation run against a different endpoint than production is measuring something else. Add output-token limits, precision, and tool-call format to the acceptance checklist alongside the model ID.

SOURCE: ARTIFICIAL ANALYSIS

PATTERN 02

Capability is rising while cost per task rises with it, at unchanged rate cards

`Muse Spark 1.2: $0.29 to $0.40 per Index task, input tokens +53% and output +36%`

`Qwen3.8-Max: 150M output tokens to run the Index against a 66M class median`

`Comparison set: Grok 4.5 (high) $0.37, GPT-5.6 Sol (medium) $0.39, Kimi K3 (max) $0.86`

Two independent releases this week got better and more expensive per unit of work without changing a published price. The mechanism is token consumption: reasoning models that think longer produce more billable output, and the effect concentrates in the long-horizon agentic evaluations where the capability gains also show up. This inverts the routing heuristic that has held for most of the year, where a cheaper rate card reliably meant a cheaper workload. Teams should re-derive per-task cost on their own traffic after every model upgrade, because the rate card no longer carries that information.

SOURCE: ARTIFICIAL ANALYSIS

PATTERN 03

Abstention is being rewarded by hallucination metrics that do not price refusal

`Muse Spark 1.2 AA-Omniscience Index 18 to 22`

`Hallucination rate 38% to 28%, but attempt rate 82% to 67% and accuracy 41% to 38%`

The headline reading of Muse Spark 1.2 is a 10-point hallucination reduction. The mechanism is that the model attempts 15 points fewer questions and its accuracy on what it does attempt fell three points. AA-Omniscience does not penalize declining to answer, so a more cautious model scores better without knowing more. This is not a criticism of the benchmark, which is measuring what it says it measures, but it is a warning about how the number travels. Any enterprise using hallucination-rate improvements to justify deployment in a high-stakes workflow should ask for the attempt rate alongside it, because a model that refuses more is a different operational product than a model that is more accurate.

SOURCE: ARTIFICIAL ANALYSIS

PATTERN 04

Revenue-gated open weights have institutionalized into their own category

`Artificial Analysis added an 'Open Weights (Commercial Use Restricted)' chart category`

`LFM Open License v1.0: no commercial use above $10M revenue, automatic termination on breach`

`Kimi K3 (W31): commercial hosting above $20M trailing revenue requires a separate agreement`

The most telling artifact this week is not a license but a chart label. Artificial Analysis introduced a distinct category defined as weights available with commercial use limited, typically requiring a paid license — which means the pattern is now frequent enough that an independent tracker needed a bucket for it. The practical consequence for architecture is that 'open weights' has stopped being a single procurement path and has become at least three: unrestricted permissive licenses, revenue-gated licenses with a negotiation path, and revenue-gated licenses without one. Liquid's is the third kind, which is the most restrictive form yet seen at this scale.

SOURCE: ARTIFICIAL ANALYSIS; LFM OPEN LICENSE V1.0

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.

ENDPOINT ACCURACY INDEX (NEW) MOVE 01

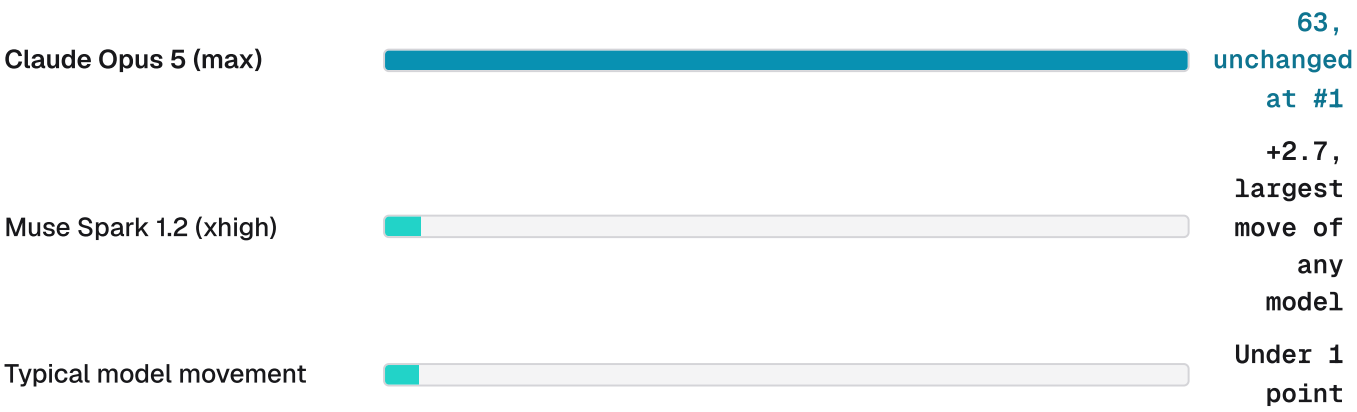
First independent measurement of how much reference accuracy each serverless endpoint preserves for identical open weights

gpt-oss-120b, worst endpoint vs reference (BFCL-500)	22% vs 37%
GLM-5.2, most restrictive endpoints (HLE-250)	≤50% of reference
DeepSeek V4 Pro, majority of endpoints	At reference parity
DeepSeek first-party endpoint	Slightly above reference

SOURCE: ARTIFICIAL ANALYSIS

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V4.1.1 MOVE 02

Grader models unified under GPT-5.6 Luna for HLE, AA-LCR and AA-Omniscience; most models moved under a point

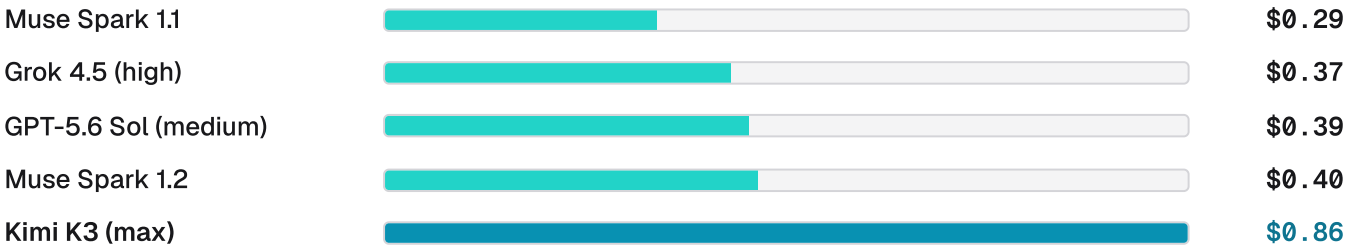


SOURCE: ARTIFICIAL ANALYSIS

COST PER INTELLIGENCE INDEX TASK

MOVE 03

Muse Spark 1.2 rose ~38% over 1.1 at unchanged list pricing, placing it mid-pack rather than cheap

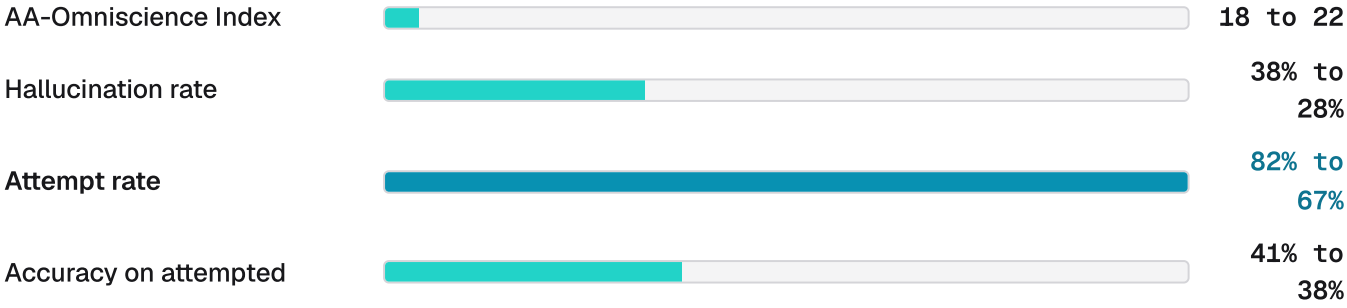


SOURCE: ARTIFICIAL ANALYSIS

AA-OMNISCIENCE (MUSE SPARK 1.1 TO 1.2)

MOVE 04

Index improved four points, but the gain is driven by the model attempting fewer questions rather than answering better



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 Aug 8, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Opus 5 — Index 63**	GPT-5.6 Sol, Claude Fable 5, Kimi K3	Unchanged at the top through the v4.1.1 patch; no closed frontier flagship shipped in-window and OpenAI delayed Astra.
Open frontier	**Kimi K3 — Index 57, revenue-tiered license**	DeepSeek V4 Pro, GLM-5.2	No new open frontier weights shipped. The Endpoint Accuracy Index now makes delivered capability depend on the serving provider, so a single leader score understates the spread.
Reasoning	**Claude Opus 5**	Qwen3.8-Max — Index 58, 1M context	Qwen3.8-Max is the week's strongest new reasoning entrant on score, and the most verbose on tokens consumed per task.
Coding	**Claude Fable 5**	GPT-5.6 Sol, Muse Spark 1.2	Epoch's MirrorCode leaderboard puts Fable 5 at 64% against GPT-5.6 Sol at 20% on long-horizon tasks; treat as a leaderboard snapshot rather than a published study.
Multimodal	**Qwen3.8-Max — text, image and video input**	Muse Spark 1.2	No dedicated multimodal flagship shipped in-window; the movement is frontier reasoning models absorbing video input rather than specialist models advancing.
Edge / small	**LFM2.5-2.6B — under 2.5 GB, ~30 tok/s on a phone**	Qwen3.5-4B, gemma-4-E2B-it	Best on-device throughput result of the week, and the only leader on this board whose license withholds commercial use from most enterprises.

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.

2026-08-05

SIGNAL 01

Meta

Contributor tier prices Muse Spark 1.2 at $0.10/$0.20 against the standard $1.25/$4.25, in exchange for permission to train on your prompts and completions

This is the clearest price yet put on enterprise data in a frontier API: roughly 92% off input and 95% off output for training rights, with rate limits applied per team rather than per key. Any team evaluating the cheap tier needs a data-governance decision before a cost decision, and the two tiers should never be mixed inside one gateway route without an explicit policy control.

SOURCE: META DEVELOPER PRICING AND RATE LIMITS DOCUMENTATION

2026-08-07

SIGNAL 02

OpenAI

Astra development partially paused and launch delayed after evaluations could not rule out a Critical cybersecurity capability level

A published capability framework has now imposed a shipping delay at the leading lab, which converts safety frameworks from disclosure documents into schedule risk that customers must model. Any roadmap contingent on a named unreleased frontier model now carries a failure mode that is neither technical nor commercial.

SOURCE: OPENAI; THE DECODER

2026-08-06

SIGNAL 03

OpenAI

The same model name now maps to different checkpoints by surface — ChatGPT chat on August Sol/Luna variants, Codex and ChatGPT Work on the July ones

Internal evaluations tied to a model name no longer transfer across products from the same vendor. Governance teams that approved 'GPT-5.6 Sol' for a use case need to record which surface and which checkpoint was tested, or the approval record is ambiguous.

SOURCE: OPENAI DEPLOYMENT SAFETY HUB

2026-08-04

SIGNAL 04

Liquid AI

Release page markets LFM2.5-2.6B as deployable 'without restrictions' while the LICENSE in the same repository withholds commercial use above $10M revenue

Procurement cannot rely on vendor release pages for license classification, because this contradiction shipped inside a single artifact on a single day. Add a step that diffs the marketing claim against the LICENSE file before any open-weight model enters evaluation.

SOURCE: LIQUID AI BLOG; LFM OPEN LICENSE V1.0

2026-08-04

SIGNAL 05

Artificial Analysis

Open-weights charts now carry a distinct 'Open Weights (Commercial Use Restricted)' category for weights whose commercial use requires a paid license

An independent tracker creating a permanent category is the strongest available evidence that revenue-gated open weights are a durable market structure rather than a run of individual vendor choices. Model shortlists should carry license class as a first-class field alongside score and price.

SOURCE: ARTIFICIAL ANALYSIS

WATCHLIST

On the radar next.

5 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.

BY AUG 31

WATCH 01

Astra's final Preparedness Framework rating and revised launch date

OpenAI reported that it cannot rule out Critical, which is not a rating. The final assessment decides whether this is the first genuine top-level capability flag or a delay narrated in safety language.

BY AUG 15

WATCH 02

Qwen-Max class open weights and the Qwen3.8-27B companion

Announced as coming 'next week' on Aug 3 with no license disclosed. Given three consecutive revenue-gated open releases, the license text will matter more than the weights.

BY SEPT 30

WATCH 03

Endpoint Accuracy Index coverage expanding beyond the launch set

Artificial Analysis named Kimi K3 as next. Broader coverage is what turns this from an interesting finding into a usable procurement instrument for open-weight shortlists.

BY OCT 31

WATCH 04

The first Gemini release under Kavukcuoglu

The flagship model is unreleased against a planned June launch and the team was reorganized on Aug 5. The next release is the first evidence of whether separating research from product execution changed shipping cadence.

ONGOING

WATCH 05

Whether cost-per-task figures get republished under Index v4.1.1

The cost-per-task comparison set was published under the prior methodology one day before the patch. If those figures are not restated, the most useful economic comparison in the industry is versioned against a superseded index.

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 16

Week 32 of 2026 · August 8, 2026

WEB

brianletort.ai/industry/model

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