THE MODEL PULSE ISSUE 18 · WEEK 34 OF 2026

THE BIG READ

OpenAI cut Sol by more than a fifth for three months while an unknown lab shipped an MIT frontier — the same week the industry conceded that vendor-reported is what the top of the board looks like now

KEY SIGNALS THIS PERIOD

1	3	3	4
Models added to the tree	Vendors shipped frontier-class	Architectural patterns crossed multiple vendors	Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

August 22, 2026

Issue 18 · 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.

Two things happened in the software layer this week that only make sense read together. On Thursday Ornith AI, a lab nobody had heard of a fortnight ago, published Ornith-1.5 under the MIT license — a 397B mixture-of-experts flagship with 35B active parameters, a 35B dense sibling, and a 9B distilled variant, benchmarked on the lab's own harness at 89.7% on Terminal-Bench 2.1 and 61.4% on DeepSWE. The next day OpenAI announced a token price cut of more than 20% on GPT-5.6 Sol, held for three months. One vendor said the top of the coding leaderboard is now open, released under a permissive license, and available to anyone. The other vendor said the current price of frontier reasoning tokens is negotiable and it is willing to say so out loud, for a fixed period, in writing.

The honest reading is that neither the leaderboard claim nor the price claim is settled. Ornith AI's Terminal-Bench figure comes from the lab's own harness, which has not been audited by any third party, and no independent tracker — Artificial Analysis, Vals, LMSYS — had ranked Ornith-1.5 by publish time. This publication is adding the model to the tree because the artifact is real and downloadable, and it is recording it at medium placement confidence with a notable field saying so plainly. If Artificial Analysis or Vals reproduces the score inside the next four weeks, the classification is trivial to lift. If they do not, that itself is the finding, and it belongs in a procurement document rather than a headline.

OpenAI's move is the more consequential one for buyers this quarter. A promotional 20%-plus cut on the flagship reasoning model, published as a dated three-month change, is the same behaviour Google put on the Gemini 3.7 Flash pricing page last week: cheap inference sold with an expiry. This publication's precision rule is that a price with a published end date is a promotional variable, not the durable rate — the durable rate is whatever the vendor's willing to write on a contract that spans the reversal. Every agent business case whose payback crosses November should model the post-promo price, and every vendor cost

comparison built on this week's per-token headline should carry the date attached.

A third release rounds out the layer without changing it. DeepSeek published DeepSeek-V4-Flash-Vision-Exp on Friday, an experimental API-only vision-capable variant of its Flash tier with no weights and no full-quality Pro comparator, aimed at multimodal harness developers testing the API surface. This publication is not adding it to the tree — models.yaml does not have a pattern for experimental API-only rows, and creating one for a variant with no weights would set a precedent the taxonomy would then have to defend against every future API-only preview. The Pulse notes it and moves on.

The absent number is the Artificial Analysis Intelligence Index. No new AA Index release landed in-window, so Grok 4.6, GPT-5.6 Sol Max and Claude Opus 5 are still separated by two points of a measurement last refreshed on August 6. Coverage will keep quoting positions on the board this week; none of them changed. And CyberGym remains where GLM-5.3 and DeepSeek-V4 Pro 0813 left it last week — vendor-reported, unreproduced, and cited interchangeably in trade press.

One finding on the served-versus-downloadable split from last week worth pulling forward: Z.ai has still not published GLM-5.3 weights to Hugging Face. The two-week window announced with the model closes at the end of this issue's coverage period. If nothing lands in the next seven days, the benchmark-first-weights-later pattern this publication flagged in W33 gets one more instance and moves toward being a durable vendor behaviour rather than a one-off. The tree keeps GLM-5.3 recorded as gated until weights are confirmed.

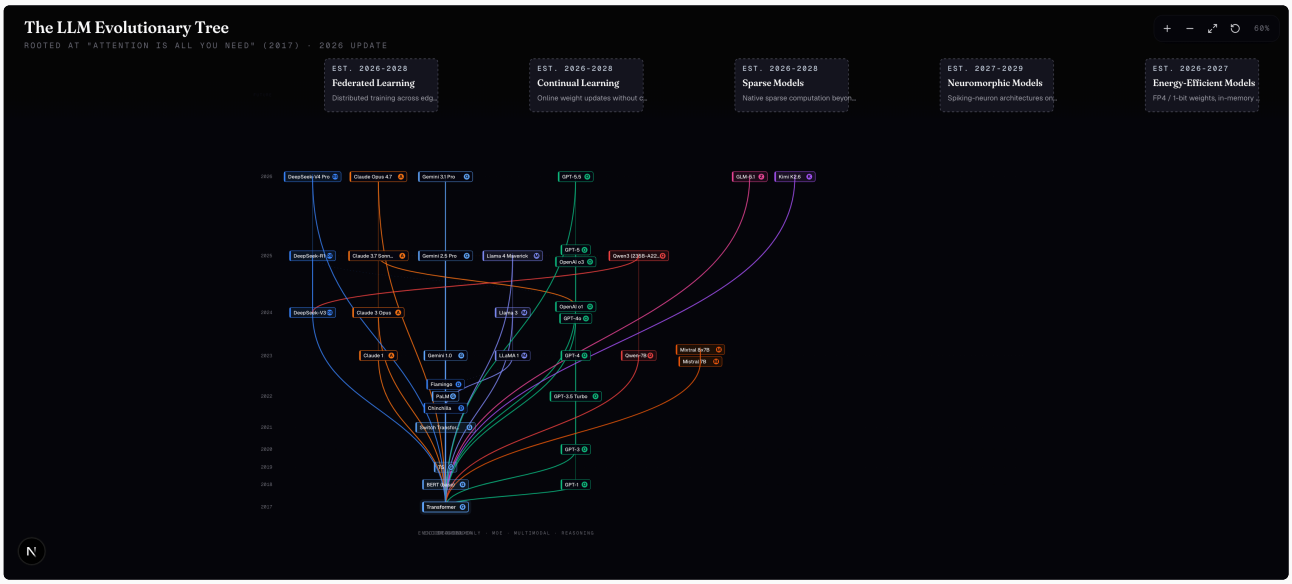
THE BAR

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.

1 model row added and 0 updated. Ornith-1.5-397B enters as a medium-confidence, open-weights MoE entry with every headline benchmark labelled vendor-reported and awaiting independent reproduction. DeepSeek-V4-Flash-Vision-Exp is deliberately kept out of the tree because models.yaml has no pattern for experimental API-only rows and one variant is a poor precedent to set.



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

ADDED (1)

ornith-1-5-397b

UPDATED (0)

No updates to existing rows this period.

The Ornith-1.5 entry carries the smallest evidence footprint the tree has admitted for a frontier claim this year: one aggregated secondary source and vendor-run harness scores. It is recorded because the weights are downloadable and MIT-licensed, which makes the artifact real regardless of the score's provenance. Reclassify to placement_confidence: high once an independent tracker reproduces at least one headline number. If the two-week Z.ai window on GLM-5.3 weights closes without a Hugging Face release, next issue will discuss whether to move GLM-5.3 off gated on the grounds that a promised-but-unshipped weight is a different classification than a shipped one.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-08-21

RELEASE 01

OpenAI

FRONTIER

REASONING

GPT-5.6 Sol (August price recut)

Token price on the flagship Sol tier cut by more than 20% for three months, disclosed as a dated promotional window rather than a durable rate change

The most economically consequential model event of the week and the one most likely to be misread as a permanent cut. OpenAI published the end date and the reversal at the same time as the reduction, which follows the pattern this publication logged with Gemini 3.7 Flash: cheap inference sold with a published expiry. Buyers should model any agent business case whose payback crosses the November reversal against the post-promo rate rather than the promotional one, and finance leaders should note that a three-month window is short enough to sit inside a single quarterly board cycle — which is likely the point. The capability did not move; only the meter did.

SOURCE: OPENAI GPT-5.6 ANNOUNCEMENT PAGE

2026-08-21

RELEASE 02

DeepSeek AI

SPECIALIST

MULTIMODAL

DeepSeek-V4-Flash-Vision-Exp

API-only experimental vision-capable variant of the Flash tier — no weights, no full Pro comparator, positioned at multimodal harness developers

Recorded on the Flash line rather than added as a new frontier row because models.yaml has no pattern for API-only experimental variants and creating one for a preview would set a precedent that fills the tree with previews. The interesting signal is what it signals about DeepSeek's release calendar: a vision experiment shipped as API-only inside two weeks of the silent V4 Pro 0813 endpoint swap suggests the vendor is iterating on the served surface faster than on the weights it publishes. Teams building multimodal harnesses should test but avoid pinning production behaviour to the experimental endpoint.

SOURCE: DEEPSEEK API DOCUMENTATION

OPEN WEIGHTS

Open-frontier and open-source drops.

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

2026-08-19

RELEASE 01

Ornith AI

OPEN FRONTIER

MOE

Ornith-1.5

397B/35B active MoE flagship (plus 35B dense and 9B distilled siblings) under MIT license, with vendor-run harness scores of 89.7% on Terminal-Bench 2.1 and 61.4% on DeepSWE

The evidence weight on this one is unusually thin and the license unusually clean, which is exactly the combination that makes it hard to file. MIT with no user gate, no naming rule and no acceptable use policy is materially more permissive than the Llama license and slightly more permissive than Apache 2.0 in effect. Every headline number is vendor-reported on a vendor-authored harness, and no independent tracker had listed Ornith-1.5 by publish time — so buyers should treat the score with the same care as any vendor benchmark and treat the artifact as real for licensing decisions but unproven for capability decisions. The 9B distilled sibling is the more immediately useful piece for teams evaluating local-first agentic tooling.

SOURCE: RUNTIMEWIRE AGGREGATION OF ORNITH AI ANNOUNCEMENT AND REPOSITORY

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

Frontier price cuts are being sold with a written expiry date attached

`GPT-5.6 Sol -20% for 3 months` `Gemini 3.7 Flash Jan 1 doubling` `DeepSeek peak/off-peak`

OpenAI made the pattern explicit this week: a headline cut of more than 20% on the flagship reasoning model, held for three months, with the reversal date published at the moment of the cut. That is the third dated promotional inference rate a major lab has shipped in five weeks. The read is that top-tier inference pricing has moved from list-price competition to promotional periods that fit inside a customer's quarterly planning window — which is easier to close deals against and much harder to build a durable margin on. Enterprise buyers should be writing model contracts with rate protection through the next reversal date and asking every vendor for one, because promo windows have become the durable form of the disclosure.

SOURCE: OPENAI GPT-5.6 ANNOUNCEMENT, GOOGLE CLOUD PRICING HISTORY

PATTERN 02

The evidence bar the tree is being asked to admit at got noticeably lower

`Ornith-1.5 vendor Terminal-Bench` `GLM-5.3 promised weights` `DeepSeek 0813 silent swap`

Three of the last two weeks' most-cited coding scores — Ornith-1.5, GLM-5.3, DeepSeek-V4 Pro 0813 — share the same evidence structure: the vendor ran the harness, the score sits at or near the top of the board, and no independent lab has reproduced it. That is not fraud, and none of these vendors dispute the labelling. It is a structural change in how a top-of-board claim now enters public awareness, and it means a procurement document quoting the ranking without the version, the harness and the vendor-reported label is measuring hype rather than capability. Treat every leaderboard row this quarter as provisional until an independent tracker reproduces the score, and treat 'the leader on X' phrasing as unverified where the leader's harness is the one measuring it.

SOURCE: ORNITH AI, Z.AI, DEEPSEEK OBSERVED RELEASE PATTERN

PATTERN 03

API-only experimental tiers as a shipping-surface for multimodal work

`DeepSeek-V4-Flash-Vision-Exp` `Gemini prior experimental Flash previews` `OpenAI Ultrafast preview`

DeepSeek followed Google's pattern of shipping vision or reasoning previews as an API-only tier separate from the weights track. This publication is choosing not to add these entries to the taxonomy as new frontier rows because the pattern is not durable inside a model card — an experimental endpoint can be retracted or renamed without a release note, and models.yaml would end up with a fossil layer of preview names. What the pattern does signal is that the served surface has become where the vendor iterates on modality, while the weight release track is where architecture is disclosed. Buyers building multimodal harnesses should assume the API surface will move faster than any published weight and design their evals to accommodate that.

SOURCE: DEEPSEEK API DOCUMENTATION, GOOGLE CLOUD GEMINI EXPERIMENTAL HISTORY

BENCHMARK MOVES

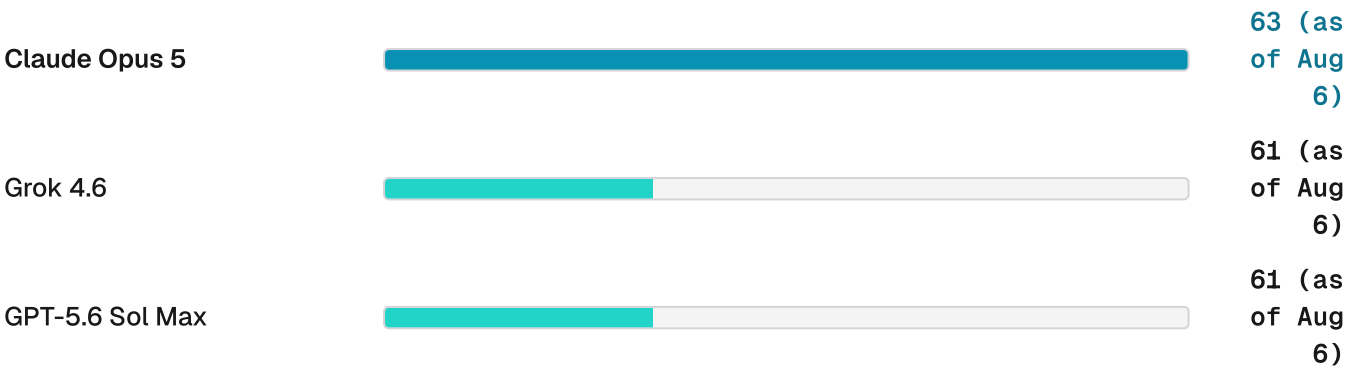
Where the leaderboard moved.

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

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX

MOVE 01

No new Index release in-window. The board is unchanged from the August 6 v4.1.1 refresh, and any coverage quoting new Index positions this week is quoting the prior refresh under a fresh headline.



SOURCE: ARTIFICIAL ANALYSIS INDEX V4.1.1 RELEASE HISTORY

TERMINAL-BENCH 2.1 (VENDOR-REPORTED ROWS)

MOVE 02

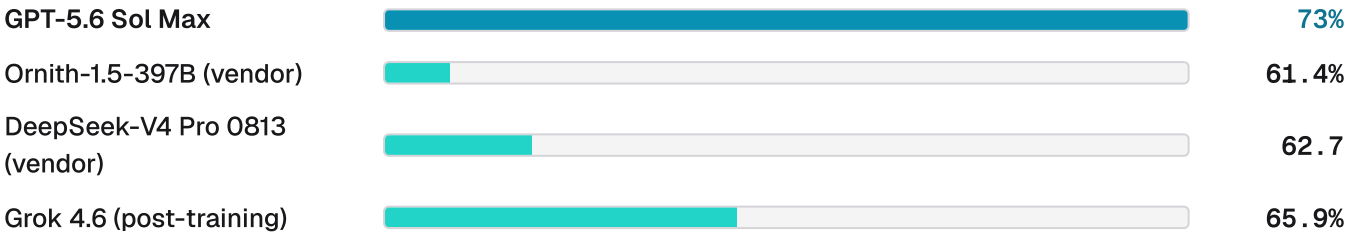
Ornith AI publishes 89.7% on its own harness, entering the top of the board on unaudited numbers. The comparability question is whether the vendor's harness measures the same thing the Anthropic-authored public evaluator measures, and this publication is not yet convinced it does.

Claude Fable 5 (v2.1)	88.0
DeepSeek-V4 Pro 0813 (v2.1, vendor)	87.9
Ornith-1.5-397B (vendor harness)	89.7
GLM-5.3 (v3.0, vendor)	top open-source claim (figure undisclosed)

SOURCE: ORNITH AI HARNESS, Z.AI RELEASE NOTES, DEEPSEEK OPENLLMSTACK AGGREGATION

DEEPSWE (SOFTWARE ENGINEERING, VENDOR ROWS) MOVE 03

Ornith-1.5 publishes 61.4% on the vendor harness, slotting between the closed-frontier ceiling and the prior open-source line. No independent reproduction attached.



SOURCE: ORNITH AI RELEASE NOTES, PRIOR PULSE ROWS

LMSYS ARENA, TEXT LISTINGS MOVE 04

Grok 4.6 continues to hold third position from last week; no new frontier listings landed in-window. Ornith-1.5 has not appeared on Arena, which is consistent with its lack of an independent tracker read.

Claude Opus 5	Top
GPT-5.6 Sol Max	Second
Grok 4.6	Third

SOURCE: LMSYS CHATBOT ARENA LEADERBOARD, SNAPSHOT AT PUBLISH

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Aug 22, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Opus 5**	Grok 4.6	Unchanged on capability from W33's read: Opus 5 holds the top of a board that did not refresh this week. The material change is on price, not on position — Sol's three-month promotional cut narrows the cost gap that separated Grok 4.6 from Sol Max on comparable tasks. Decide the tier on cost per completed task inside the promo window, and re-decide it in November.
Open frontier	**DeepSeek-V4 Pro 0813**	Ornith-1.5-397B	The challenger enters on license terms, not on independently verified capability. Ornith AI ships under MIT with 35B active MoE parameters and a 9B distilled sibling; DeepSeek's served build is still not identical to its downloadable weights. The pick this quarter depends on whether the buyer weights license permissiveness (Ornith) or served-endpoint reliability with a paid API (DeepSeek).
Reasoning	**Claude Opus 5**	GPT-5.6 Sol Max	No capability change this week. The efficiency read from W33 stands — Opus 5 remains materially more expensive per completed long-horizon task than Grok 4.6 or Sol Max — and Sol's promotional cut applies to token price, not to turn efficiency, so the durable efficiency ranking is unchanged.
Coding	**Claude Fable 5 Max**	GPT-5.6 Sol Max	Vendor-reported Terminal-Bench 2.1 numbers from Ornith and DeepSeek are not directly comparable to the audited Fable and Sol numbers because the harnesses differ. Fable 5 Max still leads CursorBench v3.2 and FrontierCode v1.1 Extended; Sol Max still leads DeepSWE v1.1. Ornith and DeepSeek's own results are on the board but not scored against them here until an independent lab reproduces one.

TIER	LEADER	CHALLENGER	READ
Multimodal	**Gemini 3.7 Flash**	Muse Spark 1.2	Unchanged. DeepSeek's V4-Flash-Vision-Exp is API-only and experimental, so it does not enter the row. Gemini 3.7 Flash's January 1 doubling still applies and this publication is still counting the row on introductory pricing.
Edge / small	**Muse Glimmer**	Ornith-1.5-9B (distilled)	Ornith AI's 9B distilled variant is the first potentially frontier-adjacent 9B open-weight release under MIT this quarter. It sits below Glimmer on Intelligence Index reads (once Glimmer is reproduced independently on the same harness) but likely above on Terminal-style tasks per vendor scores. Reclassify when a third-party benchmark lands.

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.

2026-08-21

SIGNAL 01

OpenAI

Announced a token price reduction of more than 20% on the GPT-5.6 Sol tier, effective for three months from the announcement date, with the reversal date and return-to-list rate published at launch

This is the first frontier US lab to explicitly time-box a price cut on its flagship reasoning model, and the shape matches the Gemini and DeepSeek pattern this publication has been tracking. Enterprises negotiating enterprise agreements this quarter should ask for rate protection through and beyond the November reversal, because the cut is real for the promo period and the return-to-list is equally real after.

SOURCE: OPENAI GPT-5.6 ANNOUNCEMENT PAGE

2026-08-18

SIGNAL 02

OpenAI

Announced ChatGPT ads to launch in 31 EU countries with a formal rollout date of August 24, plus a separate ChatGPT for Teens tier, positioning consumer surfaces as a distinct product line from the enterprise API

Two consumer moves that add up to a distribution story rather than a capability story. Advertising on the free tier is the first material step toward funding consumer inference through a channel other than subscription revenue, and the Teens tier is a compliance and safety product carved out ahead of the more likely EU AI Act enforcement wave. Enterprise buyers should treat neither as relevant to their contracts and note that OpenAI's product surface has visibly split into consumer and enterprise SKUs faster than any competitor's has.

SOURCE: OPENAI BLOG AND PRESS COVERAGE

2026-08-19

SIGNAL 03

Ornith AI

First public model release: Ornith-1.5 family under MIT license across a 397B/35B active MoE flagship, a 35B dense sibling, and a 9B distilled variant, published with vendor-run harness scores at or near the top of the board

The evidence structure is the tell rather than the score. A new lab shipping an MIT-licensed frontier claim inside a week of its first observed public footprint sets a very specific procurement problem: the license is clean enough to matter for legal review, and the capability claim is thin enough that no serious infrastructure decision should ride on it until an independent tracker reproduces at least one headline number. Treat the weights as usable for evaluation and the scores as unproven.

SOURCE: RUNTIMEWIRE AGGREGATION OF THE ORNITH AI RELEASE

2026-08-21

SIGNAL 04

DeepSeek AI

Published DeepSeek-V4-Flash-Vision-Exp as an API-only experimental tier with no accompanying weights and no full Pro-tier vision comparator

The Pulse is reading this as a shipping-surface signal rather than a capability signal: DeepSeek is iterating multimodal work on the API surface rather than on the weight track, which is consistent with the silent 0813 endpoint swap two weeks ago. Multimodal harness developers may test it, but no production system should depend on an experimental endpoint that ships without an announcement or a stability commitment. Recorded but not added to the tree.

SOURCE: DEEPSEEK API DOCUMENTATION

2026-08-22

SIGNAL 05

Z.ai

GLM-5.3 weights still not on Hugging Face at publication; Z.ai's own two-week promise window from the August 14 announcement closes at the end of this issue's coverage period

This publication does not move a model out of gated on promises. If nothing lands in the next seven days, next issue will discuss whether the benchmark-first-weights-later pattern rises to the level of a durable vendor behaviour worth naming, and whether models.yaml should distinguish 'weights promised' from 'weights shipped' as separate rather than transitional states.

SOURCE: Z.AI HUGGING FACE REPOSITORY STATUS AT PUBLISH, W33 PULSE

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 26

WATCH 01

NVIDIA Q2 FY2027 earnings, data-center revenue, and any Rubin Ultra memory statement

The single largest scheduled disclosure in the quarter, and the earliest chance for NVIDIA to confirm or deny the 8-stack HBM4E de-spec reporting that has been carried unresolved for two weeks. Any official memory spec commentary would resolve W33's second-noise call directly, and would be the first primary source on Rubin Ultra configuration rather than a rumor-based one.

AUG 22 - SEP 5

WATCH 02

Independent reproduction of Ornith-1.5 headline scores on Artificial Analysis or Vals

The bar this publication is applying is that an MIT-licensed frontier claim from an unknown lab needs at least one independent leaderboard read before the tree lifts placement confidence. Two weeks is a reasonable window for a public tracker to test a downloadable model, and either a confirmation or a contradicting result is decision-relevant for procurement.

AUG 29

WATCH 03

Z.ai two-week window on GLM-5.3 weights closes

The announced window ends on or around this date. Whether the weights land is the difference between a shipped open-source model and an announcement, and this publication will name the pattern as durable if a second frontier open-source release ships weights on a similar delay this quarter.

NOV 21

WATCH 04

OpenAI GPT-5.6 Sol promotional pricing window closes

The dated reversal on the 20%-plus cut. Watch whether OpenAI extends, replaces or reverts on the exact date — an extension would soften the durable-rate reading of the cut, a reversion would confirm it. Any enterprise contract signed against the promotional rate needs a plan for that Monday.

SEP-OCT

WATCH 05

Artificial Analysis Intelligence Index v4.2 or v4.1.2 refresh

The Index has not refreshed since v4.1.1 on August 6, which means all Index positions currently in trade press are two-plus weeks stale. The next refresh will be the first opportunity to see whether Ornith-1.5, GLM-5.3 or the DeepSeek 0813 build enter the ranked board on independent scoring.

SEP-OCT

WATCH 06

Anthropic in-house silicon commentary and shipping timeline signal

Anthropic hired Amir Salek (a TPU program veteran) this week for in-house silicon per secondary reporting; the Pulse does not carry the hire as a model event but flags any subsequent official confirmation or Trainium-independence commentary from Anthropic as the next relevant data point. A durable capability shift shows up as vendor guidance to customers, not as a personnel move.

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 18

Week 34 of 2026 · August 22, 2026

WEB

brianletort.ai/industry/model

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