THE MODEL PULSE ISSUE 21 · WEEK 37 OF 2026

THE BIG READ

The serving layer changed underneath models already deployed, and this week's new checkpoints lead on cost and footprint rather than capability

KEY SIGNALS THIS PERIOD

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

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

September 12, 2026

Issue 21 · 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.

W37 produced three material changes to models already deployed and two genuinely new open-weight releases, neither of which leads on capability. DeepSeek retired V4-Flash on September 10 and announced that from 04:00 UTC on September 14 every deepseek-v4-pro request routes to V4.1-Flash and bills at V4.1-Flash rates until a V4.1-Pro exists. GitHub deprecated a model across chat, inline edits, ask and agent modes on September 10, with a replacement that administrators may have to enable themselves. Artificial Analysis revised its Intelligence Index on September 7, four days after the previous revision, rescaling the composite so that last week's 61 and this week's 53 are not the same measurement.

Read together, those three are one event: the identity of the model behind a pinned endpoint is now a vendor-controlled variable, and so is the scale of the number you used to justify choosing it. None of the three required a new model to be released. All three change what an already-deployed system does.

The genuine releases of the week reinforce the point rather than contradicting it. DeepSeek published an MIT-licensed 552B multimodal Mixture-of-Experts model whose headline property is a key-value cache of 890 bytes per token, roughly a quarter of its predecessor's, which is a serving-economics claim rather than a capability claim. OpenAI made a full-duplex voice model generally available at $0.05 per minute and split the conversational layer from the reasoning layer, so a voice agent is now two rate cards that can change independently. Nex-AGI published a 1.6 trillion-parameter open checkpoint whose own reference deployment requires sixteen H200-class accelerators, which makes it nominally open and practically gated by hardware. W36 already argued that model diligence had moved off the model card and onto the harness, so the direction is not new. What W37 adds is the other three surfaces and a date for each: an endpoint reassigned to a different model on September 14, a managed service whose residency and retention posture is documented rather than inferred, and an index whose version number now changes the score. The defensible W37 read is that

harness diligence has become endpoint, residency and index-version diligence as well, and that all four moved inside one week.

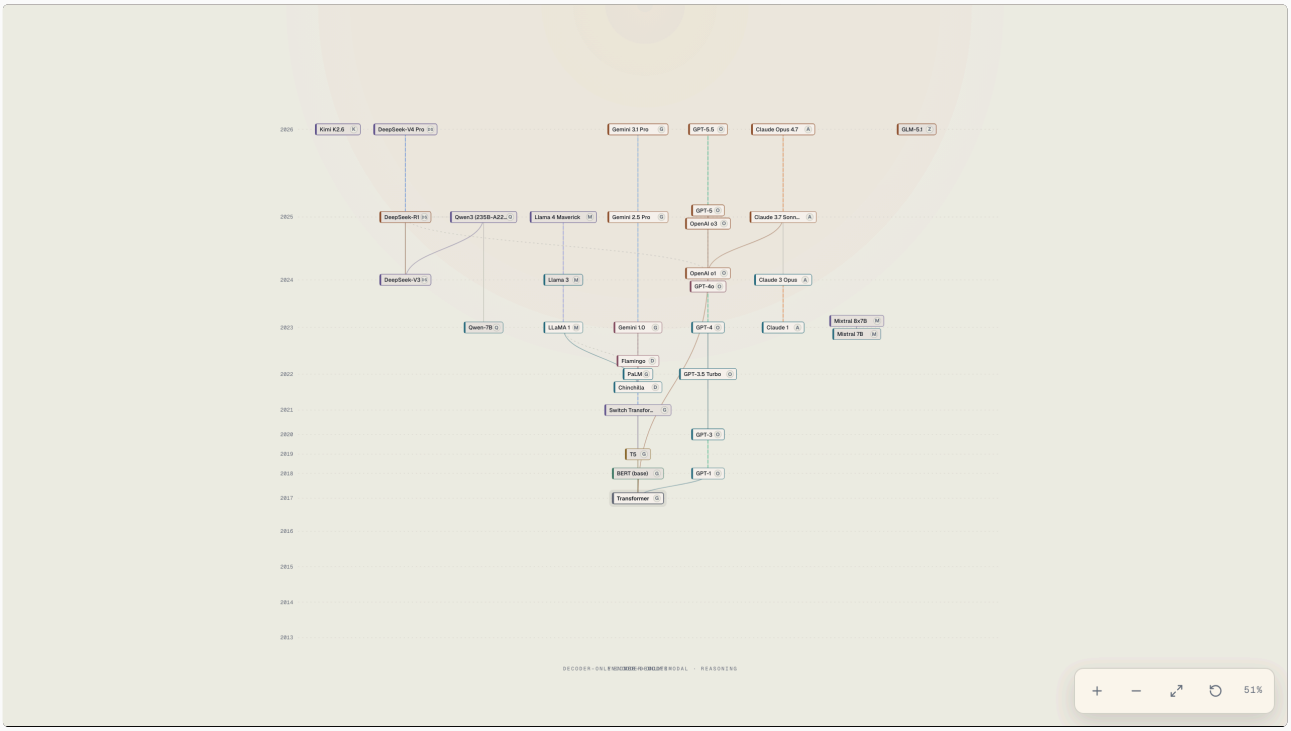
THE BAR

A model release matters when it changes one of three things: which workload runs on-prem, which tier of license is viable for a deployment, or which capability ceiling the closed frontier enforces. Single-vendor releases that change none of these are noise.

TREE DELTA

What changed in the tree.

Three rows added and two updated. The additions are the week's two open-weight release events; the updates are both in the DeepSeek V4 line and are the more consequential half of the delta, because they change what an already-deployed endpoint serves.



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

ADDED (3)

- deepseek-v4-1-flash
- nex-n2-5-max
- nex-n2-5-mini

UPDATED (2)

- deepseek-v4-pro
- deepseek-v4-flash

deepseek-v4-1-flash enters as an MIT-licensed 552B multimodal Mixture-of-Experts at 1M context, and nex-n2-5-max and nex-n2-5-mini enter as the two Nex-N2.5 checkpoints whose weights were actually published. The 397B Nex-N2.5 Pro is deliberately not added: its weights were listed as coming soon inside the window and it was reachable only through hosted access, so there is nothing yet to place. On the update side, deepseek-v4-flash is retired as of September 10 and deepseek-v4-pro records that its endpoint routes to a different, smaller model from September 14 with no change required by the caller. Read the three additions as cost and footprint claims rather than capability claims, because that is what both releases lead with.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-09-10

RELEASE 01

OpenAI

FRONTIER

MULTIMODAL

GPT-Live-1

Full-duplex voice generally available at $0.05 per minute, with reasoning delegated to a separately billed model

The model listens and speaks simultaneously and hands reasoning and tool calls to a backend text model the developer selects and pays for separately. That makes a voice agent a two-line bill whose latency and cost can be tuned independently, and it also means the deployment now depends on two rate cards that can change without notice. OpenAI reports a 30-percentage-point improvement on a full-duplex benchmark over its predecessor, which is vendor-reported.

SOURCE: OPENAI

2026-09-03

RELEASE 02

OpenAI

FRONTIER

REASONING

GPT-6 Astra

Re-measured mid-window at an identical rounded score to Claude Fable 5.1 but 57% lower measured cost per task

Artificial Analysis Intelligence Index v4.3, published September 7, puts Astra and Fable 5.1 level at 53 while measuring cost per index task at $3.26 against $7.63. The rank is an artifact of a benchmark revised twice in four days and should not be read as model progress. The cost separation is the durable procurement finding, and it survives the rescaling because it is a measured spend figure rather than a composite score.

SOURCE: ARTIFICIAL ANALYSIS

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.

2026-09-10

RELEASE 01

DeepSeek

OPEN FRONTIER

MOE

DeepSeek-V4.1-Flash

MIT-licensed 552B multimodal Mixture-of-Experts with a 1M context and an 890-byte-per-token cache

The defining number is not a benchmark but a memory figure: DeepSeek reports a global key-value cache of 890 bytes per token, roughly a quarter of its predecessor's, which is what makes a million-token context economically serveable rather than merely supported. It exceeds the larger V4-Pro it replaces on two of DeepSeek's own reported benchmarks, DeepSWE v1.1 at 74.2 resolved against 62.7 and Terminal-Bench 2.1 at 90.6 Pass@1 against 87.9, both vendor-run and neither independently reproduced. The weights are published under MIT, so the serving economics are available to anyone with the hardware. Note a disclosed inconsistency: the model card states a 552B backbone while repository metadata lists 763B, and DeepSeek does not explain the difference.

SOURCE: DEEPSEEK MODEL CARD

2026-09-08

RELEASE 02

Nex-AGI

OPEN FRONTIER

MOE

Nex-N2.5 Max

1.6 trillion-parameter open checkpoint whose own reference deployment needs sixteen H200-class accelerators

This is the clearest current example of an open weight that is nominally available and practically gated: the vendor's reference serving configuration specifies two nodes of eight H200 GPUs with tensor and expert parallelism at sixteen. Read it as a post-training achievement rather than a new foundation, because independent coverage reports it is built on DeepSeek-V4-Pro-Base, and its 1.6 trillion total and 49 billion active parameters match that model exactly. For an organisation that already holds multi-node H200 capacity it is a real self-hosting option at trillion-parameter scale. For everyone else the 35B sibling is the only member of the family that can actually be piloted.

SOURCE: NEX-AGI MODEL CARD

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

Cache residency, not parameter count, is the long-context lever

`DeepSeek-V4.1-Flash` `GLM 5.3 under vLLM Hybrid HiSparse`

Two independent pieces of work in the same week attacked long context through memory policy rather than model size. DeepSeek reports a global key-value cache of 890 bytes per token, roughly a quarter of its predecessor's, at a 1,048,576-token context. vLLM published a residency policy that keeps cache on the accelerator while capacity allows and offloads all but indexer-selected tokens to host memory under pressure, which it states lets GLM 5.3 run its full million-token context on a single eight-GPU node where that was previously impossible. Both reframe long context as a bytes-per-token budgeting problem, which is the number to put in a capacity plan.

SOURCE: VLLM, DEEPSEEK MODEL CARD

PATTERN 02

The conversational layer separates from the reasoning layer

`GPT-Live-1`

A full-duplex voice model priced per minute, delegating reasoning and tool calls to a separately selected and separately billed text model, splits what used to be one procurement into two. The upside is real: cheap models can handle scheduling turns while expensive ones handle escalations, and latency can be tuned where it is felt. The cost is that the deployment's economics now sit on two rate cards, either of which can be repriced or rerouted by its vendor, and this week supplied an example of exactly that happening elsewhere in the stack.

SOURCE: OPENAI

PATTERN 03

The serving floor decides who can actually use an open model

`Nex-N2.5 Max` `Nex-N2.5 mini` `DeepSeek-V4.1-Flash`

Open weights published this week span a two-order-of-magnitude range in what it takes to run them, and the licence tells you almost nothing about that. A 1.6 trillion-parameter checkpoint arrives with a sixteen-accelerator reference configuration, a 35B sibling arrives as the pilotable member of the same family, and a 552B model arrives with a cache figure specifically low enough to make its full context affordable. The practical reading is that openness and accessibility have decoupled, so an open-weight strategy now needs a hardware floor stated alongside every candidate model.

SOURCE: NEX-AGI MODEL CARD, DEEPSEEK MODEL CARD

PATTERN 04

Inference silicon routes around high-bandwidth memory rather than competing for it

d-Matrix Raptor | Positron Asimov

Two inference-silicon events in the same week share one design thesis: avoid the constrained memory and advanced-packaging supply chain instead of bidding inside it. One uses logic on three-dimensional DRAM and enters through an established rack interconnect ecosystem, with a tape-out targeted before the end of 2026 and rack availability in the fourth quarter of 2027. The other uses standard low-power memory to reach 288 GB to 2,304 GB per chip, taping out at the end of 2026 for production in the second half of 2027. Neither has silicon; both are bets that memory supply rather than compute is the binding constraint on cost per token.

SOURCE: D-MATRIX, POSITRON AI

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 V4.3

MOVE 01

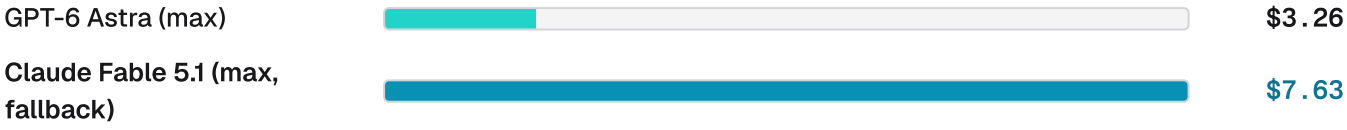
Rescaled twice in four days, so last week's 61 and this week's 53 are different measurements and any reading of a decline is wrong

SOURCE: ARTIFICIAL ANALYSIS

COST PER INTELLIGENCE INDEX TASK (V4.3)

MOVE 02

The durable finding of the week is price rather than rank: a 57% separation at an identical rounded score

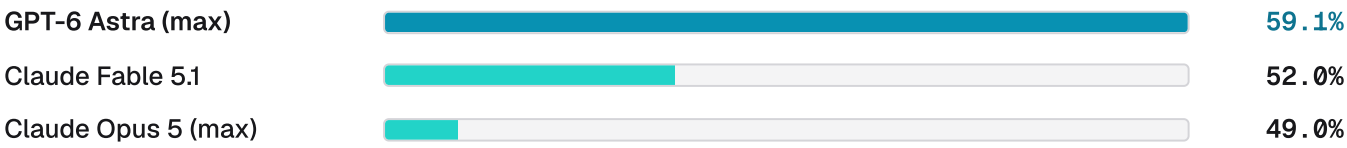


SOURCE: ARTIFICIAL ANALYSIS

TERMINAL-BENCH V4.0

MOVE 03

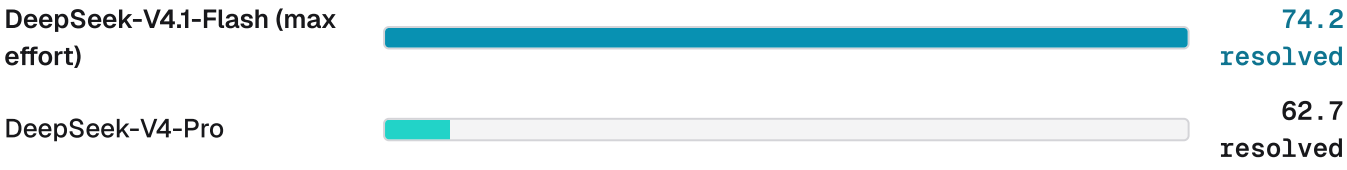
Upgraded from v2.1 inside the same index revision, so prior Terminal-Bench figures do not carry forward into comparisons



SOURCE: ARTIFICIAL ANALYSIS

DEEPSWE V1.1 (VENDOR-REPORTED) MOVE 04

A smaller successor beats the larger flagship it replaces on the vendor's own harness, which is the stated basis for retiring the flagship endpoint



SOURCE: DEEPSEEK MODEL CARD

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Sep 12, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Fable 5.1**	GPT-6 Astra	Level at 53 on Intelligence Index v4.3, but the tie is a rounding artifact on a scale revised twice in four days; treat the $3.26 against $7.63 cost-per-task spread as the real separation.
Open frontier	**DeepSeek-V4.1-Flash**	Nex-N2.5 Max	MIT weights at 552B with an 890-byte-per-token cache against a 1.6T checkpoint needing sixteen H200-class accelerators; the licence is no longer the constraint, the serving floor is. Independent coverage of Nex-AGI's materials, not a Nex-AGI page, reports that the challenger is post-trained from DeepSeek's own V4-Pro base; its parameter shape matches exactly, which makes this tier closer to one lineage at two price points than to two competitors.
Reasoning	**Claude Fable 5.1**	GPT-6 Astra	No reasoning-specific release landed in the window, so the standing order is unchanged and rests on an index whose composition changed underneath it on September 7.
Coding	**GPT-6 Astra**	Claude Fable 5.1	Astra leads Terminal-Bench v4.0 at 59.1% against 52.0%, on a benchmark version introduced this week, so this ordering has no comparable prior-week figure.
Multimodal	**Gemini 3.8 Flash**	DeepSeek-V4.1-Flash	Native video input remains the Flash family's distinguishing modality rather than a single model's, and DeepSeek's new multimodal open weight is the first credible open alternative at frontier scale.
Edge / small	**Nex-N2.5 mini**	Desert Ant on-device model set	A 35B open checkpoint against eighteen task-specific models that run entirely on device with no per-inference bill below a stated free tier; both are pilotable without data centre capacity.

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.

2026-09-10

SIGNAL 01

DeepSeek

Flagship endpoint rerouted to a smaller model, and its predecessor retired the same day

From 04:00 UTC on September 14 every deepseek-v4-pro request serves V4.1-Flash at V4.1-Flash rates until a V4.1-Pro ships. Any team with the flagship pinned in code gets a different model and a different bill without changing a line, which makes endpoint identity a term to pin contractually rather than a constant.

SOURCE: DEEPSEEK API CHANGE LOG

2026-09-10

SIGNAL 02

GitHub

A model deprecated the same day across chat, inline edits, ask and agent modes

The replacement may need enterprise administrator enablement, so an organisation running scheduled or unattended agent work can have that work change behaviour without shipping a release of its own. Version the model your automations depend on and test scheduled agents against model changes you do not control.

SOURCE: GITHUB CHANGELOG

2026-09-07

SIGNAL 03

Artificial Analysis

Second Intelligence Index revision in four days, with private-test weighting raised to 45%

The index replaced a banking benchmark with a 657-task held-out business-workflow set and upgraded Terminal-Bench two major versions. Any model-selection business case quoting an Artificial Analysis score from before September 7 must now carry its index version or it is not defensible in review.

SOURCE: ARTIFICIAL ANALYSIS

2026-09-10

SIGNAL 04

OpenAI

Agents API ships in public beta with US-only residency and no zero data retention

OpenAI's own documentation states the managed harness supports US data residency only and does not support zero data retention, and that choosing a self-hosted sandbox does not make it eligible. That disqualifies it for regulated workloads today regardless of capability, which makes this a developer-velocity decision rather than a production one.

SOURCE: OPENAI DEVELOPER DOCS

2026-09-09

SIGNAL 05

Anthropic

Lab discloses that its own agentic transcript review missed evaluation incidents

Anthropic states its original review spanned roughly 141,000 transcripts and that because the scan relied on agentic search it missed a set of transcripts that also had internet access, surfacing a fourth incident only while assembling material for an external evaluator. It is a direct argument for deterministic coverage checks on any agentic evidence review used in audit.

SOURCE: ANTHROPIC

2026-09-10

SIGNAL 06

European Commission

EU cybersecurity agency granted hands-on access to two frontier models for vulnerability testing

Access followed one model's release by about a week and the other by more than five months from announcement, with roughly three months of safeguard negotiation. The gap shows the arrangement is negotiated bilaterally rather than compelled, so regulator visibility currently scales with each vendor's willingness rather than a uniform standard.

SOURCE: REUTERS

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.

SEP 14

WATCH 01

DeepSeek flagship reroute takes effect

Every deepseek-v4-pro request begins serving V4.1-Flash at V4.1-Flash rates; re-baseline any evaluation or cost model pinned to the flagship endpoint.

SEP 15-17

WATCH 02

AI Infra Summit silicon status

Marvell's 102 Tbps Ethernet switch and 256-lane scale-up switch were announced with no silicon status, node, power, or availability date; the show is where those should appear.

SEP 21-23

WATCH 03

First multi-vendor 1600ZR interoperability demonstrations

The Implementation Agreement published September 9 has no shipping module against it yet, so interop results are the first real test of whether 1.6T coherent is procurable multi-vendor.

LATER IN 2026

WATCH 04

vLLM v0.30 and the cache-residency policy

Hybrid HiSparse is not in any release and needs a pinned checkout today, so the single-node million-token capacity claim only enters a plan when v0.30 ships.

DEC 31, 2026

WATCH 05

Gemini 3.8 Flash introductory pricing expiry

Current Flash-tier task economics rest on introductory provider pricing; the defensible 2027 run rate is whatever rate survives the reversal.

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 21

Week 37 of 2026 · September 12, 2026

WEB

brianletort.ai/industry/model

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