THE MODEL PULSE ISSUE 15 · WEEK 31 OF 2026

THE BIG READ

Open weights finally became deployable artifacts — and the closed stack answered with post-training and rate cards, not new flagships

KEY SIGNALS THIS PERIOD

1	6	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 1, 2026

Issue 15 · 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.

Moonshot's Jul 27 Kimi K3 weight drop converts last week's promissory open-frontier watch item into a procurement fact: a 2.8T MoE with 104B active parameters, 896 experts selecting 16, hybrid 69 KDA + 24 Gated MLA layers, 1,048,576 context, and MoonViT-V2 — under a revenue-tiered Kimi K3 License that triggers a separate commercial deal for MaaS above $20M trailing-twelve-month revenue. Self-host evaluation can now start on artifacts; license diligence is the gate, not model quality.

No closed frontier model shipped. The closed stack competed on price and post-training instead. OpenAI cut GPT-5.6 Luna 80% to $0.20/$1.20 and Terra 20% to $2/$12 on Jul 30 while holding Sol at $5/$30 and replacing Priority Processing with Fast mode at 2x token price for up to ~2.5x speed. DeepSeek put V4-Flash-0731 into public beta on Jul 31 on the same 284B/13B architecture and unchanged $0.14/$0.28 pricing; Artificial Analysis independently scored it +10 Intelligence Index points to 50. Thinking Machines released Inkling-Small the same day as the Luna cut — 276B/12B active, Apache 2.0, independent AA Index 40 versus Inkling's 41 — with a context-window conflict to resolve (vendor card up to 1M; AA article lists 256K).

The week's largest vendor-claimed capability jump did not need new weights either: OpenAI reported ARC-AGI-3 public-set scores moving from 13.3% to 38.3% on GPT-5.6 Sol by retaining reasoning between tool calls and compacting context instead of truncating, with roughly 6x fewer output tokens. That is a harness result, not an official leaderboard reprint — and no qualifying LMArena, SWE-bench, or ARC official board moves landed in-window. Gemini 3.5 Pro missed its Jul 31 GA prediction. MiniMax H3 promised open weights and had not shipped them as of Aug 1. Net/net: route on measured Index and task cost, treat open as license-plus-weights, and re-test agents against memory policy before buying a bigger model.

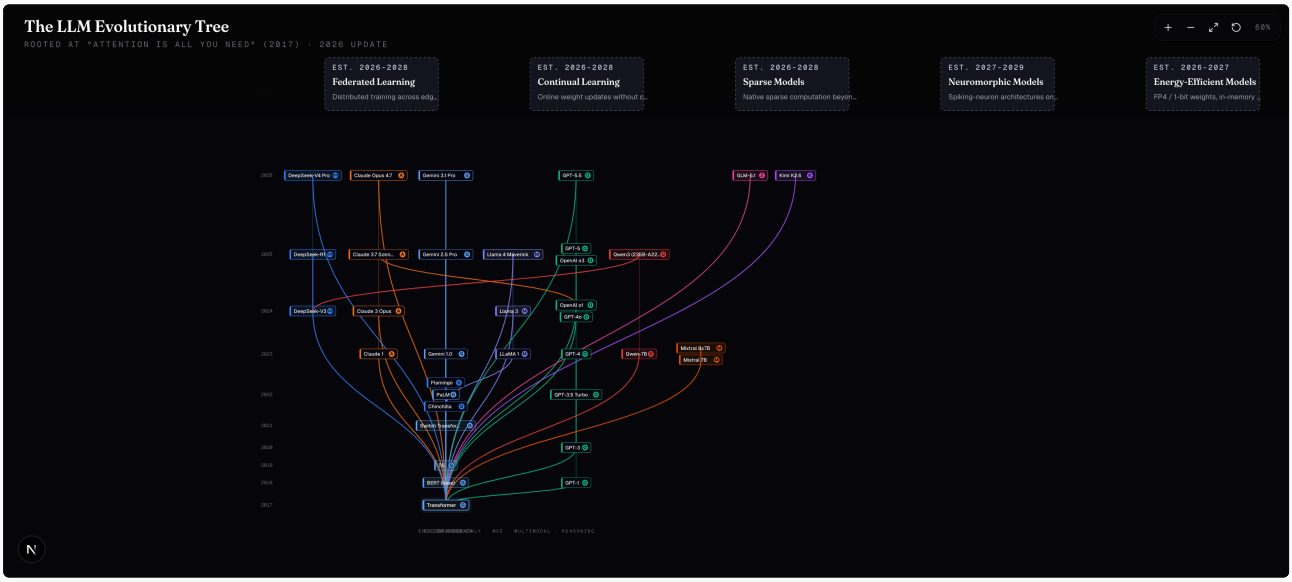
THE BAR

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.

Inkling-Small joins the tree as the strongest US open-weights efficiency result to date, reaching an independent Intelligence Index of 40 against Inkling's 41 with under a third of the parameters. Kimi K3 is reclassified from closed to open weights on the July 27 artifacts and license text, with its active-parameter count corrected from a community estimate to a confirmed 104B. DeepSeek V4 Flash updates for the 0731 re-post-train checkpoint that gained 10 independent Index points on unchanged architecture.



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

ADDED (1)

inkling-small

UPDATED (2)

kimi-k3

deepseek-v4-flash

Kimi K3 is an update rather than an addition: the row has existed since W29, and this week's material change is openness, license terms, and a corrected active-parameter count of 104B against the earlier community-estimated ~50B. Qwen3.7 Flash is held out of the tree pending a first-party model card, and MiniMax H3 is deferred until its promised weights actually ship. Both are covered in the cards.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-07-30

RELEASE 01

OpenAI

FRONTIER

REASONING

GPT-5.6 Luna / Terra price cut and Sol Fast mode

Luna falls 80% to $0.20/$1.20, Terra 20% to $2/$12, Sol holds $5/$30, and Fast mode sells ~2.5x speed at 2x token price

Re-baseline high-volume routing immediately: Luna undercuts most Western cheap tiers on sticker price while Sol remains the premium band against Opus 5. Treat Fast mode as a purchasable latency tier rather than a separate model, and instrument per-tier completion rates before moving fleet volume on list price alone. The AI Stack Weekly owns the fixed-workload spread calculation; this Pulse stays on the architecture and routing implication.

SOURCE: OPENAI; UNITE.AI; VENTUREBEAT; THE DECODER

2026-07-31

RELEASE 02

DeepSeek

FRONTIER MOE

DeepSeek V4-Flash-0731

Same 284B/13B MoE and $0.14/$0.28 pricing, re-post-trained into public beta; independent AA Index jumps 10 points to 50

Drop this into existing Flash routes as a cheap-tier agent upgrade without a size or price change — but do not treat it as V4-Pro GA, and pin the harness when comparing Terminal-Bench (vendor-claimed 82.7 max effort versus Artificial Analysis 79%). Architecture and model ID stay deepseek-v4-flash; Pro, app, and web remain unchanged pending the promised official V4-Pro.

SOURCE: DEEPSEEK API CHANGELOG; ARTIFICIAL ANALYSIS (INDEPENDENT)

2026-07-29

RELEASE 03

OpenAI

REASONING

AGENTIC

GPT-5.6 Sol harness memory on ARC-AGI-3

Vendor-reported public-set score triples from 13.3% to 38.3% via retained reasoning plus context compaction on the same model

Treat the ARC-AGI-3 jump as vendor-claimed on a single harness, not an official leaderboard reprint: OpenAI reports roughly 6x fewer output tokens when compaction replaces rolling truncation on the Responses API, and no qualifying LMArena or SWE-bench official moves landed in-window. Agent Techniques Weekly owns the method and the reproduction caveat.

SOURCE: OPENAI (VENDOR-REPORTED HARNESS RESULT)

2026-07-27

RELEASE 04

Alibaba / Qwen

SPECIALIST

MULTIMODAL

Qwen3.7 Flash

Quiet low-cost multimodal API at reported $0.03/$0.13 under 32K and 1M context — no lab blog, no weights, no public benchmark table

Useful as a fleet-cost floor candidate for multimodal agents, but treat specs and pricing as secondary until a first-party model card is pinned. Sources are OpenRouter listing dates and technical digests citing a QwenCloud changelog; no qwenlm.github.io launch post was found in-window.

SOURCE: EESEL.AI; BENCHLM (SECONDARY; NO PRIMARY LAB ANNOUNCEMENT)

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.

2026-07-27

RELEASE 01

Moonshot AI

OPEN FRONTIER

MOE

Kimi K3

Full weights ship: 2.8T MoE, 104B active, 896/16 experts, hybrid 69 KDA + 24 Gated MLA, 1,048,576 context, revenue-tiered license

Start the self-host evaluation now — GitHub and Hugging Face artifacts landed Jul 27 — but send the Kimi K3 License to counsel before any commercial hosting plan: MaaS above $20M aggregate trailing-12-month revenue needs a separate deal, with attribution triggers above 100M MAU or $20M monthly revenue. Independent AA Index reference remains 57; vendor-claimed table includes GPQA Diamond 93.5 and DeepSWE 67.5. Hosted API stays $3/$15 (cached input $0.30).

SOURCE: MOONSHOTAI/KIMI-K3 GITHUB; HUGGING FACE; LICENSE TEXT; UNITE.AI

2026-07-30

RELEASE 02

Thinking Machines Lab

OPEN FRONTIER

MOE

Inkling-Small

276B/12B-active multimodal MoE under Apache 2.0; independent AA Index 40 vs Inkling 41 at under a third the parameters

Strongest US open-weights efficiency play this week for teams that cannot absorb K3-class cluster scale: BF16 needs ~600 GB aggregate VRAM, NVFP4 ~180 GB, with vLLM v0.26.0 shipping same-week Inkling family support. Flag the context conflict honestly — vendor model card says up to 1M; Artificial Analysis lists 256K in its article blurb — and use Index 40 as the independent baseline rather than vendor-claimed SWE-bench Verified 80.2% alone.

SOURCE: THINKING MACHINES LAB; ARTIFICIAL ANALYSIS (INDEPENDENT INDEX 40); VENTUREBEAT

2026-07-31

RELEASE 03

MiniMax

SPECIALIST

MULTIMODAL

MiniMax H3

Multimodal video model launches API-first with open weights promised in coming days — still not on Hugging Face or GitHub as of Aug 1

Treat current state as API-only until the repo appears: unified text/image/video/audio understanding to video with native stereo, up to 15s at 2K, with vendor claims that 2K per-second price is under one-third of mainstream models. Do not build a self-host plan around the promise; the W30 K3 lesson applies again.

SOURCE: MINIMAX BLOG; SCMP; SECONDARY WEIGHTS-NOT-YET STATUS

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 and harness policy move scores without new weights

`DeepSeek V4-Flash-0731: +10 AA Index on unchanged 284B/13B`

`OpenAI ARC-AGI-3: 13.3% to 38.3% via retained reasoning + compaction`

`OpenAI Jul 29 serving/harness efficiency post (~20% serving cost, >15% token-gen efficiency)`

Two independent vendors moved measured capability this week without shipping new parameter counts: DeepSeek re-post-trained Flash in place, and OpenAI changed Responses API memory settings on the same Sol weights. Architects should treat reasoning retention, context compaction, and post-training checkpoints as first-class evaluation dimensions alongside the model ID. Buyers who only diff weight releases will miss the largest week-over-week gains.

SOURCE: DEEPSEEK; ARTIFICIAL ANALYSIS; OPENAI

PATTERN 02

Revenue-tiered open licenses turn legal review into the deployment gate

`Kimi K3 License: >$20M trailing-12-month MaaS commercial trigger`

`Kimi K3 attribution triggers above 100M MAU or $20M monthly revenue`

`Inkling-Small: clean Apache 2.0 contrast case`

Downloadable weights no longer imply a single open-source diligence checklist. K3 is MIT-like until commercial hosting crosses the revenue trigger, at which point counsel must negotiate; Inkling-Small remains Apache 2.0 with no such trigger. Procurement should classify each open release by license economics before sizing clusters, and keep Apache-clean alternatives on the shortlist when MaaS revenue is material.

SOURCE: MOONSHOTAI/KIMI-K3 LICENSE; THINKING MACHINES LAB

PATTERN 03

Serving stacks ship in the same week as the open MoE class

`vLLM v0.26.0: Inkling modeling stack, NVFP4, MTP speculative decoding`

`Inkling-Small BF16 ~600 GB / NVFP4 ~180 GB VRAM floors`

`DeepSeek-V4 routing and speculative optimizations across NVIDIA/AMD/XPU`

Production self-hosters got a same-week inference path for the new open MoE class rather than a multi-month serving lag. vLLM v0.26.0 (Jul 27) adds Inkling CUDA-graph and Hopper FA4 relative attention support alongside DeepSeek-V4 serving pushes and matured KV offloading. Capacity planners should price FP4 datapaths and KV-tiered storage into the model-release decision, not as a later optimization.

SOURCE: VLLM PROJECT GITHUB RELEASE V0.26.0; THINKING MACHINES MODEL CARD

PATTERN 04

Near-iso Intelligence at a fraction of active parameters

`Inkling-Small: AA Index 40 vs Inkling 41 at <1/3 parameters`

`Inkling-Small: 12B active vs Inkling 41B active`

`DeepSeek V4 Flash prior: also sat at AA Index 40 in the same size class`

Artificial Analysis places Inkling-Small within one Index point of full Inkling while activating 12B rather than 41B parameters, and notes no smaller open-weights model scores higher in that size class. Efficiency MoEs are now a procurement tier of their own: teams should ask whether the last Index point is worth 3x the serving footprint before defaulting to the flagship open checkpoint.

SOURCE: ARTIFICIAL ANALYSIS (INDEPENDENT)

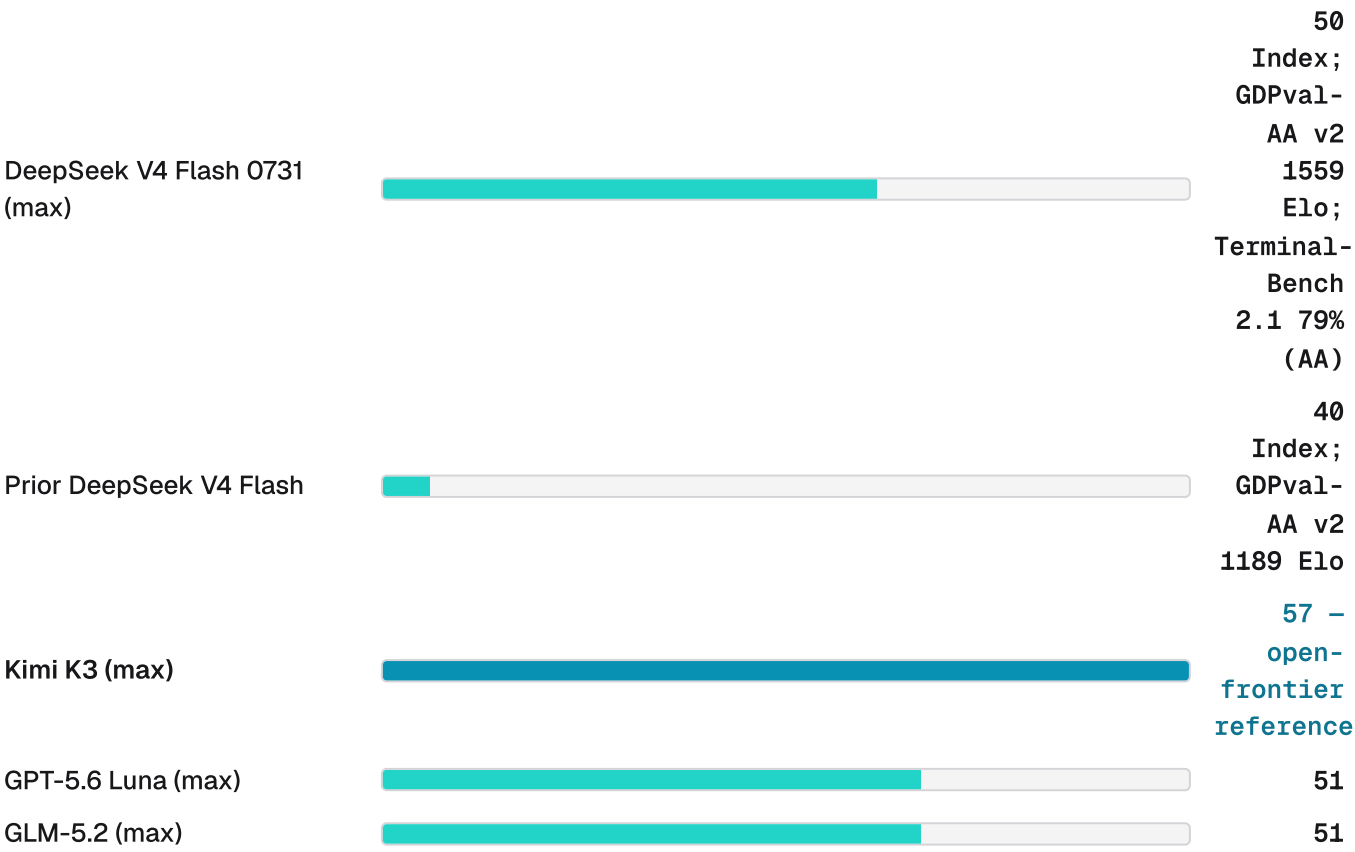
BENCHMARK MOVES

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 — DEEPSEEK V4 FLASH MOVE 01

Independent Index jumps from 40 to 50 (+10) for Flash-0731 on unchanged architecture; now ~1 point behind GPT-5.6 Luna (51) and 6 points above V4 Pro on AA's comparison

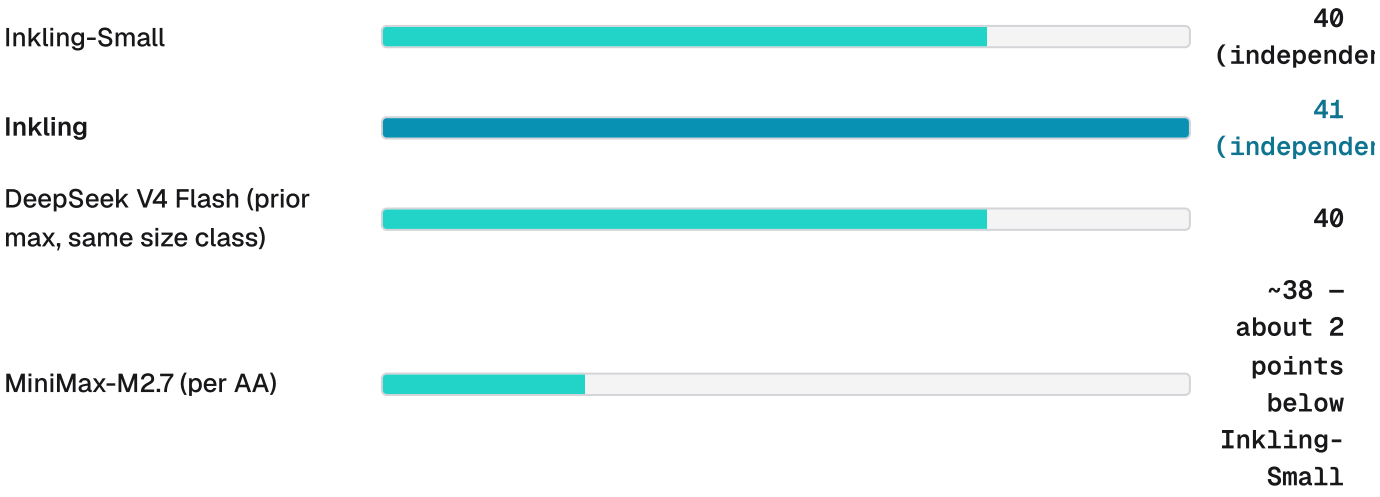


SOURCE: ARTIFICIAL ANALYSIS (INDEPENDENT)

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX — INKLING-SMALL

MOVE 02

New open US model lands at Index 40, within 1 point of Inkling (41) at less than a third the parameters

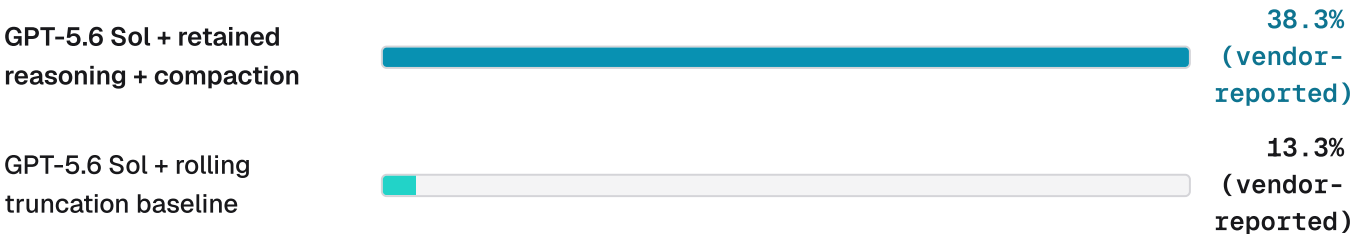


SOURCE: ARTIFICIAL ANALYSIS (INDEPENDENT)

ARC-AGI-3 PUBLIC SET (VENDOR HARNESS, NOT OFFICIAL BOARD)

MOVE 03

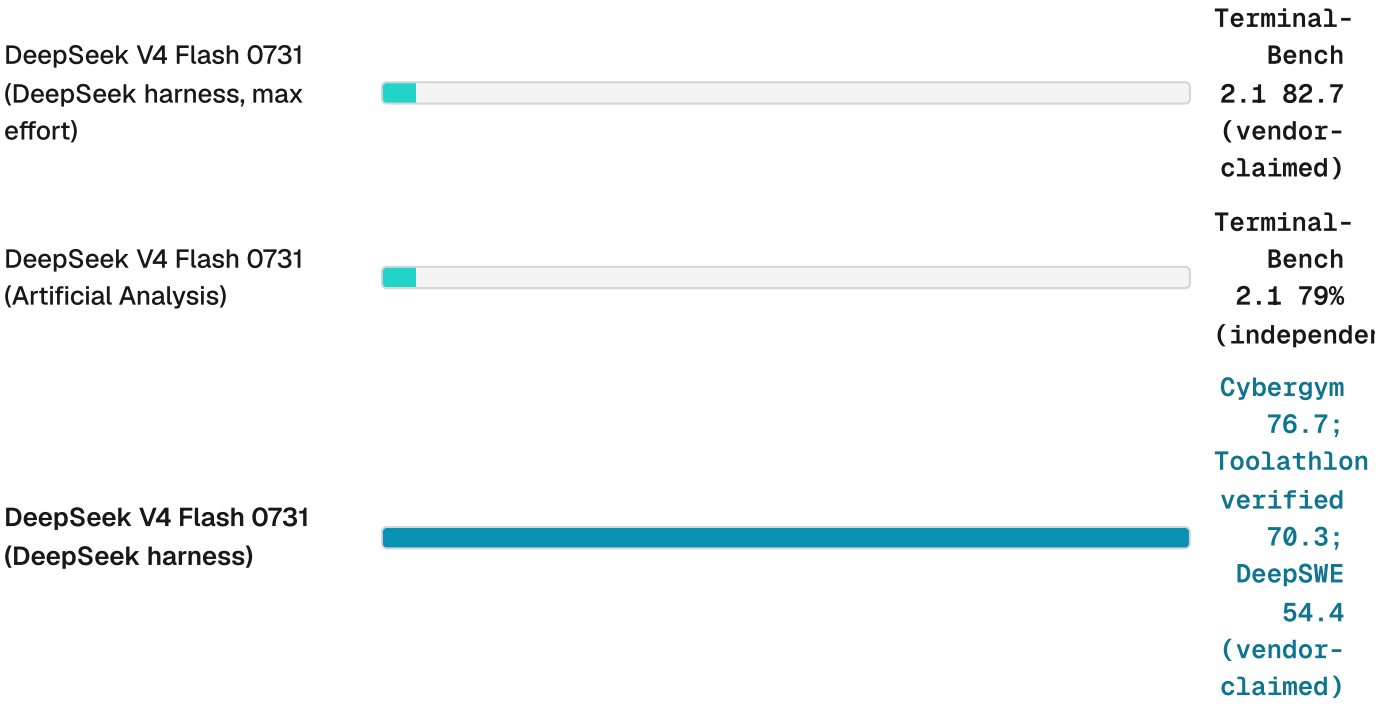
OpenAI reports GPT-5.6 Sol rising from 13.3% to 38.3% by retaining reasoning and compacting context — same model, roughly 6x fewer output tokens



SOURCE: OPENAI (VENDOR-REPORTED; NOT AN OFFICIAL ARC LEADERBOARD REPRINT)

DEEPSEEK TERMINAL-BENCH 2.1 (VENDOR HARNESS VS INDEPENDENT) MOVE 04

Vendor-claimed max-effort 82.7 versus Artificial Analysis 79% on the same Flash-0731 checkpoint — pin the harness before ranking



SOURCE: DEEPSEEK API CHANGELOG (VENDOR); ARTIFICIAL ANALYSIS (INDEPENDENT)

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Aug 1, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Fable 5**	Claude Opus 5	No new closed flagship shipped; Fable retains the absolute reference while Opus 5 remains the economic default at $5/$25.
Open frontier	**Kimi K3**	Inkling-Small	K3 takes the lead now that weights and license text are public (AA Index 57); Inkling-Small is the efficiency challenger at Index 40 under Apache 2.0.
Reasoning	**Claude Fable 5**	GPT-5.6 Sol	Sol's vendor-reported ARC-AGI-3 harness jump shows memory policy can move scores without new weights; independent cost-per-task still pending.
Coding	**Claude Opus 5**	DeepSeek V4 Flash 0731	Opus 5 holds the practical coding lead on W30 launch evidence; Flash-0731 is the cheap-tier agent upgrade at Index 50 and unchanged $0.14/$0.28.
Multimodal	**Gemini 3.6 Flash**	Kimi K3	Gemini remains the public fleet multimodal default; K3 now adds self-host control with MoonViT-V2. Gemini 3.5 Pro still has not GA'd.
Edge / small	**Inkling-Small**	Gemini 3.5 Flash-Lite	Inkling-Small brings open multimodal agents to ~180 GB NVFP4; Flash-Lite remains the hosted speed-cost floor at roughly 350 tok/s.

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.

2026-07-30

SIGNAL 01

OpenAI

GPT-5.6 Luna cut 80% to $0.20/$1.20 and Terra 20% to $2/$12; Sol held at $5/$30; Fast mode replaces Priority Processing

Closed labs are competing on completed-task economics via rate cards and latency tiers rather than new flagship weights this week. Route only latency-sensitive work into Fast mode and keep ordinary asynchronous agents on standard Luna or Terra after measuring completion quality.

SOURCE: OPENAI

2026-07-27

SIGNAL 02

Moonshot AI

Kimi K3 License ships with weights: MIT-like until MaaS revenue exceeds $20M trailing twelve months

Open-weights diligence now splits into quality evaluation and license classification. Counsel should map expected MaaS and product revenue against the trigger before any production self-host commitment; Apache-clean alternatives remain available for teams that cannot accept the commercial clause.

SOURCE: MOONSHOTAI/KIMI-K3 LICENSE; HUGGING FACE

2026-07-31

SIGNAL 03

Google DeepMind

Gemini 3.5 Pro still not generally available as the standing Jul 31 prediction deadline expires

Do not plan Q3 procurement around 3.5 Pro. Buy the callable Flash and Flash-Lite fleet now; another slip further weakens roadmap-based purchasing.

SOURCE: GOOGLE (JUL 21 PARTNER-TESTING STATUS; NO IN-WINDOW GA FOUND)

2026-07-31

SIGNAL 04

MiniMax

H3 launches with an explicit promise to open weights in the coming days; no public HF or GitHub artifacts as of Aug 1

Repeat the K3 lesson: an open-weights commitment is not a downloadable artifact. Keep H3 on the API watchlist until the repo and license appear.

SOURCE: MINIMAX BLOG; SECONDARY WEIGHTS-NOT-YET REPORTS

2026-07-27

SIGNAL 05

vLLM project

vLLM v0.26.0 ships Inkling family support, NVFP4, MTP speculative decoding, and DeepSeek-V4 serving optimizations

Serving readiness is now part of the model-release race. Self-host teams evaluating Inkling-Small or DeepSeek V4 this week should pin v0.26.0 as the baseline inference path rather than waiting for a later stack catch-up.

SOURCE: VLLM GITHUB RELEASES

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.

NEXT 7-14 DAYS
WATCH 01

MiniMax H3 downloadable weights and license

Decides whether H3 becomes a real open video-generation option or remains an API product wrapped in open-language marketing.

NEXT 30 DAYS
WATCH 02

DeepSeek V4-Pro official GA and pricing

Flash-0731 is not Pro GA; the standing surge-pricing and official Pro triggers remain incomplete until a callable Pro ID and rate card land.

NEXT 14 DAYS
WATCH 03

Inkling-Small context window: vendor 1M vs AA 256K

Procurement specs need a resolved context number; watch for an AA methodology note or Thinking Machines clarification.

BY AUG 15
WATCH 04

Independent Opus 5 cost-per-task results

Vendor benchmark leadership from W30 is still not enough; the market needs task cost, token use, and latency against Fable 5 and GPT-5.6 Sol.

NEXT 2-3 WEEKS
WATCH 05

LMArena Opus 5 Elo stabilization

No qualifying in-window official board move; secondary pages remain unstable — wait for a date-stamped stable print before re-ranking coding.

NEXT 30 DAYS
WATCH 06

Gemini 3.5 Pro callable model ID and public pricing

The Jul 31 miss is now a track-record item; only a public GA closes the W30 prediction cleanly.

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 15

Week 31 of 2026 · August 1, 2026

WEB

brianletort.ai/industry/model

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