THE MODEL PULSE ISSUE 13 · WEEK 29 OF 2026

THE BIG READ

The 'cheap Chinese open model' era ended in one launch: Kimi K3 topped an LMArena flagship board, priced itself at Claude Sonnet parity — and shipped without weights. Meanwhile the largest US-origin open-weights model actually landed, and OpenAI showed what it keeps for itself.

KEY SIGNALS THIS PERIOD

2	4	3	4
Models added to the tree	Vendors shipped frontier-class	Architectural patterns crossed multiple vendors	Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

July 18, 2026

Issue 13 · 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 Kimi K3 (Jul 16) is the week's headline and its own cautionary tale. The 2.8T-parameter MoE became the first open-weight-lab model to top an LMArena flagship board (#1 Frontend Code Arena at 1,679 preliminary, ahead of Fable 5 and GPT-5.6 Sol) and debuted #4 on the AA Intelligence Index at 57.1 — but it launched hosted-only, with weights promised by Jul 27 and no license published, so Artificial Analysis classifies it proprietary in the interim. And the price tells the strategic story: $3/$15 per M tokens, flat across the full 1M context, is roughly 3x the K2.6 tier and parity with Claude Sonnet 5. The standing procurement assumption that Chinese open models cost 10x less than US closed ones is over; budget models on measured cost-per-task, not on lineage. The week's actual open-weights event came from Thinking Machines Lab: Inkling (Jul 15), a 975B/41B-active trimodal MoE under clean Apache 2.0 with BF16 and NVFP4 checkpoints and day-0 support in every major serving stack — the largest US-origin open-weights model to date, positioned explicitly as a fine-tuning base rather than a chat flagship. And OpenAI's GPT-Red disclosure (Jul 15) marks the capability-gating precedent to watch: an internal-only red-teaming model trained at flagship post-training scale that beat human red-teamers 84% to 13% on indirect prompt injection and hardened GPT-5.6 Sol to a 0.05% direct-injection failure rate. The operational read across all three: treat 'open' as a license-and-weights fact, not a launch label; make prompt-injection robustness a quantified line item in model selection; and expect the strongest capabilities to increasingly ship as vendor-internal advantages rather than products.

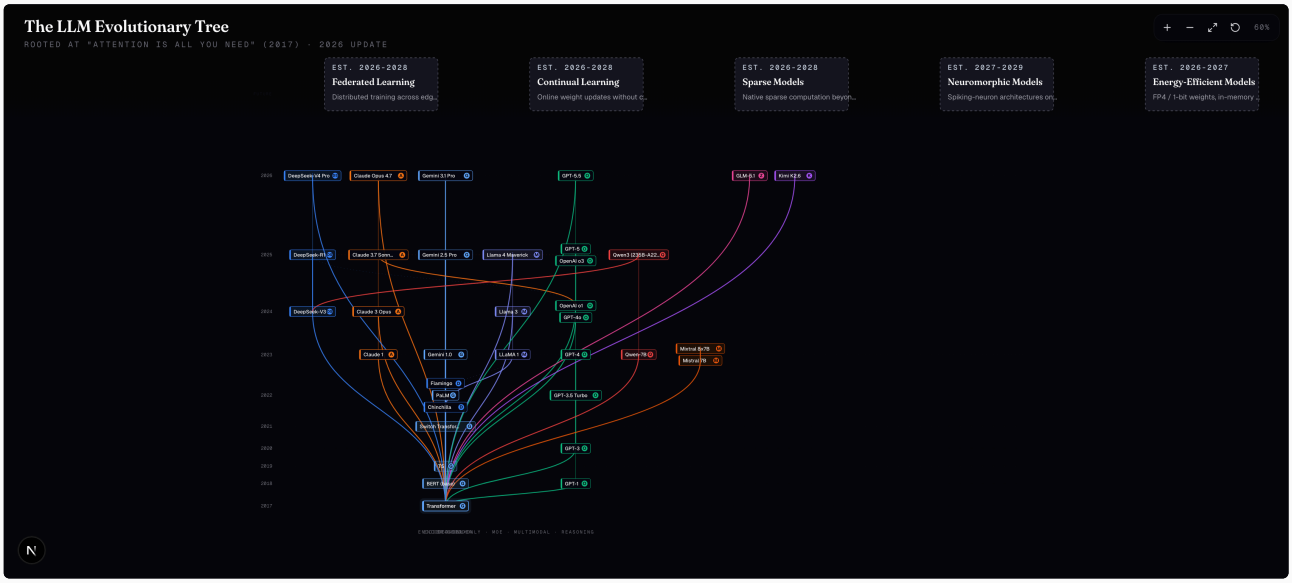
THE BAR

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.

Two rows added: kimi-k3 lands Moonshot's 2.8T MoE on the mixture-of-experts branch, recorded as closed until the promised weights and license actually ship, and inkling adds the largest US-origin open-weights model — a 975B trimodal Apache 2.0 MoE from Thinking Machines Lab.



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

ADDED (2)

kimi-k3
inkling

UPDATED (0)

No updates to existing rows this period.

GPT-Red is excluded from the tree by design — OpenAI states it will never be released, and the tree tracks deployed or deployable models. The reported Gemma 4 in-place refresh is weakly sourced (single secondary blog) and is covered under vendor signals rather than as a tree update.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-07-15

RELEASE 01

OpenAI

SPECIALIST

REASONING

GPT-Red

An internal-only self-play RL red-teaming model, trained at flagship post-training scale, that beat human red-teamers 84% to 13% on indirect prompt injection — and will never be released

Procurement should now treat prompt-injection robustness as a quantified, vendor-differentiating spec: GPT-Red hardened GPT-5.6 Sol to a 0.05% direct-injection failure rate (6x fewer than the best model four months prior), and its attack corpora have been folded into every OpenAI release since GPT-5.3 — numbers you can demand equivalents of from every vendor. The withholding itself is the second signal: frontier labs are starting to keep their strongest capabilities as internal advantages, so vendor selection increasingly buys the byproducts of models you will never see. All figures are vendor-reported; the technical preprint promised the week of Jul 20 is the verification checkpoint.

SOURCE: OPENAI (VENDOR-REPORTED; PREPRINT PROMISED WEEK OF JUL 20); LIVE ANDON LABS VENDING-AGENT COMPROMISE DEMO

2026-07-13

RELEASE 02

OpenAI via AWS

FRONTIER

REASONING

GPT-5.6 family (Sol / Terra / Luna) on Amazon Bedrock

Full family GA on Bedrock's Responses API endpoint at OpenAI first-party rates that count toward AWS commitments — with 90% prompt-cache discounts via explicit cache breakpoints

Teams with AWS enterprise commitments can now route frontier OpenAI traffic through existing spend at no rate premium — re-run the make-vs-route math on any direct OpenAI contract renewal. Mind the caveats before migrating: context is capped at 272K on Bedrock, and Sol is limited to us-east-1/us-east-2 (Terra and Luna add us-west-2), so region-sensitive workloads need the smaller tiers. The 90% prompt-cache discount with explicit breakpoints makes cache design, not list price, the dominant cost lever for agent fleets on this endpoint.

SOURCE: AWS WHAT'S NEW; BEDROCK MODEL CARD

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.

2026-07-16

RELEASE 01

Moonshot AI

FRONTIER MOE

Kimi K3

2.8T MoE (16 of 896 experts, ~50B active), 1M context, Kimi Delta Attention, MXFP4 quantization-aware training — hosted-only at launch, with weights promised by Jul 27

Hold the 'open' celebration until Jul 27: K3 launched API-only, the license is unpublished, the HF repo 404'd the day after launch, and AA classifies it proprietary in the interim — so treat it as a closed model with an open-weights option pending, and gate any self-hosting plan on the actual license text. The capability is real regardless: #1 on the LMArena Frontend Code Arena (first open-weight-lab model to top a flagship board), #4 on the AA Intelligence Index, and the best overall SWE Marathon score (42.0, vendor-reported on the KimiCode harness). Plan for heavy serving if weights do land — Moonshot recommends 64+ accelerator supernodes — and note it launched with only reasoning_effort=max, so token spend runs verbose (130M tokens on the AA eval vs a 63M peer average).

SOURCE: MOONSHOT AI; ARTIFICIAL ANALYSIS; LMARENA; SIMON WILLISON

2026-07-15

RELEASE 02

Thinking Machines Lab

OPEN FRONTIER

MOE

Inkling

The largest US-origin open-weights model: 975B/41B-active trimodal MoE under Apache 2.0, 1M context, 45T training tokens, BF16 + NVFP4 checkpoints, day-0 transformers/vLLM/SGLang/llama.cpp

If your open-model strategy assumed the frontier-scale open lane belonged to Chinese labs, revise it: Inkling is a clean Apache 2.0, US-origin base explicitly positioned for fine-tuning and domain adaptation, with serving partners (Baseten, Modal, Databricks, TogetherAI, Fireworks) live at launch and a managed post-training platform (Tinker) attached. Budget realistically — ~2TB VRAM for BF16 or ~600GB for the calibrated NVFP4 checkpoint — and watch the previewed Inkling-Small (276B/12B active) as the practical deployment tier for most teams. The day-0 NVFP4 release on GB300-trained infrastructure also signals where open-model serving economics are heading: 4-bit datapaths as the default, not an afterthought.

SOURCE: THINKING MACHINES LAB VIA HUGGING FACE; NVIDIA (GB300 NVL72 TRAINING CONFIRMATION)

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

Extreme MoE sparsity converges on ~2-4% activation at trillion scale

`Kimi K3 (16 of 896 experts, ~1.8%)` `Inkling (41B of 975B, 4.2%)` `DeepSeek V4 Pro (49B of 1.6T, ~3%)`

Three independent trillion-scale designs now activate only 2-4% of parameters per token, which means serving cost tracks memory capacity and interconnect bandwidth, not FLOPs — the constraint that matters when sizing inference fleets for these models. It also makes routing stability a first-order engineering concern (Quantile Balancing, Per-Head Muon are the current techniques), because a mis-routed expert at 1.8% activation wastes a much larger share of the useful compute. Capacity planners should price these models on HBM footprint and east-west bandwidth, and treat vendor active-parameter counts as the real sizing number.

SOURCE: MOONSHOT AI; THINKING MACHINES LAB; DEEPSEEK

PATTERN 02

4-bit-native training collapses the release-to-deployment gap

`Kimi K3 (MXFP4/MXFP8 QAT from the SFT stage)` `Inkling (calibrated NVFP4 day-0, GB300-trained)`

Both of the week's big releases trained with 4-bit precision in the loop rather than quantizing after the fact — K3 ran quantization-aware training from the SFT stage, and Inkling shipped a calibrated NVFP4 checkpoint the same day as BF16. That removes the accuracy-loss lottery that used to sit between an open-weights release and a production deployment, and it couples open models to FP4-capable datapaths. Procurement implication: hardware refresh decisions should now weight native FP4 support heavily, because the open-model ecosystem is standardizing on it at the source.

SOURCE: MOONSHOT AI; THINKING MACHINES LAB VIA HUGGING FACE

PATTERN 03

1M context becomes the default tier — with flat pricing

`Kimi K3 (KDA, flat $3/$15 across 1M)` `Inkling (1M)` `DeepSeek Sparse Attention`

Hybrid linear/sparse attention designs (K3's Kimi Delta Attention, DeepSeek Sparse Attention) have made million-token context cheap enough that Moonshot prices it flat — the same $3/$15 per M tokens at 1M as at 1K, with no long-context surcharge. Once long context stops carrying a price penalty, prompt-cache hit rate becomes the dominant cost lever for agent workloads: the difference between a well-designed cache ($0.30 cached-input on K3, 90% discounts on Bedrock GPT-5.6) and a naive one dwarfs the model-choice delta. Architect agent memory around cache breakpoints before optimizing anything else.

SOURCE: MOONSHOT AI; AWS BEDROCK DOCUMENTATION

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.

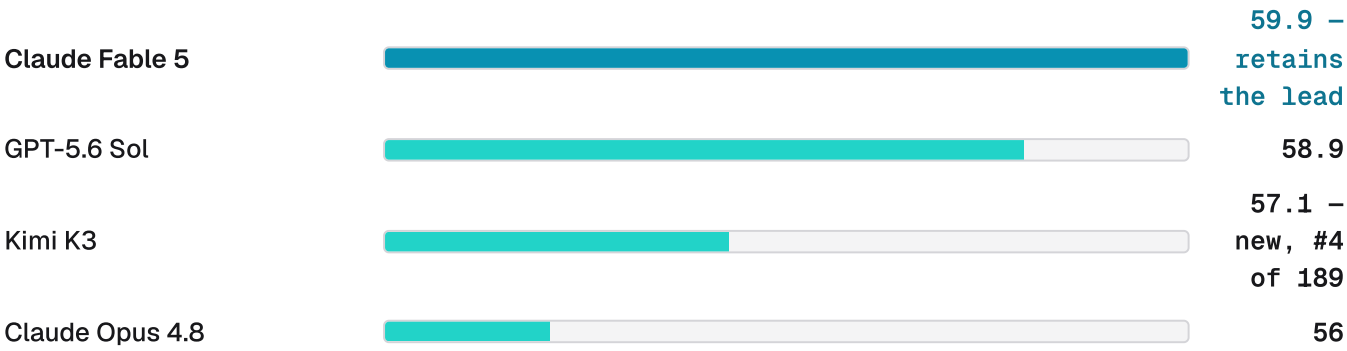
LMARENA FRONTEND CODE ARENA MOVE 01

Kimi K3 debuted at #1 with 1,679 (preliminary) — the first model from an open-weight lab to top an LMArena flagship board, 48 points clear of Claude Fable 5

SOURCE: LMARENA

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX MOVE 02

K3 entered #4 of 189 at 57.1, above Opus 4.8 — but generated 130M tokens on the eval (63M peer average) at $0.94 per Index task, and AA classifies it proprietary until weights land

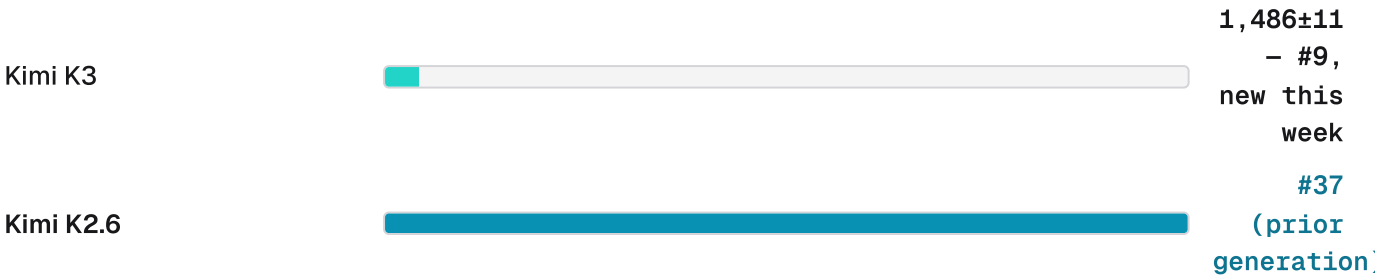


SOURCE: ARTIFICIAL ANALYSIS

LMARENA TEXT ARENA

MOVE 03

K3 entered #9 at 1,486±11 — a 28-place jump over K2.6's #37, moving Moonshot from mid-pack to the top-10 conversation tier in one generation

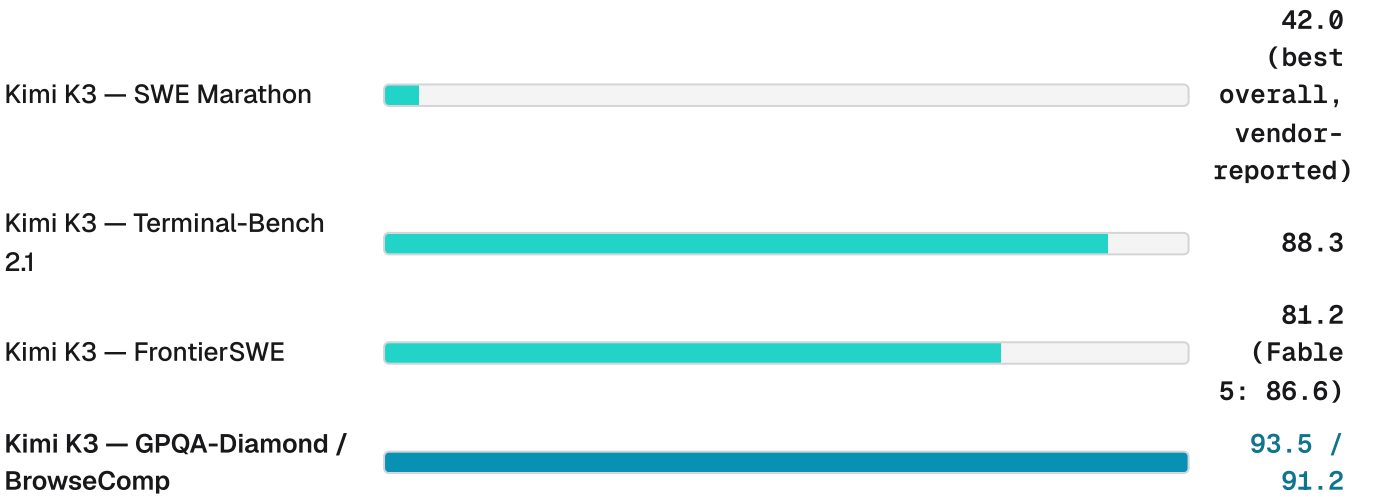


SOURCE: LMARENA

AGENTIC CODING SUITE (VENDOR-REPORTED, KIMICODE HARNESS)

MOVE 04

K3 posted the best overall SWE Marathon score and near-frontier terminal results — but every number comes from Moonshot's own harness, and W28's lesson was that harness choice can swing cost and scores 2x, so compare against Sol's Codex-harness numbers with care



SOURCE: MOONSHOT AI (VENDOR-REPORTED, KIMICODE HARNESS)

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Jul 18, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Fable 5**	GPT-5.6 Sol	Fable 5 keeps the AA Intelligence Index lead (59.9 vs 58.9) and the coding crown; Sol's counter this week is distribution (Bedrock GA at first-party rates) and hardening (0.05% direct-injection failure rate off GPT-Red) rather than raw capability.
Open frontier	**GLM-5.2**	Kimi K3	GLM-5.2 holds on the strength of last week's independent Databricks validation; K3 credibly challenges on scores (#4 AA Index, #1 Frontend Code Arena) but stays challenger until the promised weights and license actually ship Jul 27 — Inkling enters the tier as the cleanest-licensed base at this scale.
Reasoning	**Claude Fable 5**	GPT-5.6 Sol	Unchanged at the top; K3 launched with only reasoning_effort=max and verbose token behavior (130M tokens on the AA eval), so its reasoning economics are unproven until the promised low/high effort settings arrive.
Coding	**Claude Fable 5**	Kimi K3	K3 displaces GLM-5.2 as challenger: #1 Frontend Code Arena (1,679 prelim vs Fable 5's 1,631) and the best SWE Marathon score, though the agentic-suite numbers are vendor-reported on Moonshot's own harness while Fable 5 holds FrontierSWE (86.6 vs 81.2).
Multimodal	**Gemini 3.5 Flash**	Kimi K3	Gemini 3.5 Pro missed its reported Jul 17 target — the third slip; K3 enters as challenger with text+image+video input at 1M context, and Inkling's trimodal (text/image/audio) Apache 2.0 base gives builders an open multimodal foundation for the first time at this scale.
Edge / small	**Mellum2**	Inkling-Small (previewed)	Inkling-Small (276B/12B active, weights promised after testing) would reset the efficient-deployment tier if it ships as previewed; until then the Hy3

TIER	LEADER	CHALLENGER	READ
			local-quant pipeline from W28 remains the practical near-frontier-at-home path.

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.

2026-07-16

SIGNAL 01

Moonshot AI

K3 pricing resets the Chinese-lab tier: $0.30 cached / $3.00 input / $15.00 output per M tokens — input up 216% and output up 275% versus K2.6, at parity with Claude Sonnet 5

The 'Chinese open models are 10x cheaper' planning assumption is dead: Moonshot is pricing on capability, not on lane. Re-run any routing or budget model built on that assumption — the K2.x tiers remain available as the value option, but the frontier Chinese tier now costs what the US mid-frontier costs.

SOURCE: MOONSHOT AI PRICING DOCUMENTATION; SIMON WILLISON

2026-07-16

SIGNAL 02

Moonshot AI

K3 launched as 'open' with no weights: hosted-only, license unpublished until the promised Jul 27 drop, HF repo 404'd the day after launch — AA classifies it proprietary in the interim

'Open' is becoming a launch-marketing label decoupled from the license-and-weights fact. Contract and compliance teams should gate any K3 self-hosting or fine-tuning commitment on the actual license text, and treat 'weights coming' announcements as vendor roadmap, not product.

SOURCE: ARTIFICIAL ANALYSIS; SIMON WILLISON

2026-07-14

SIGNAL 03

DeepSeek

Hard retirement of the deepseek-chat / deepseek-reasoner aliases lands Jul 24, alongside the first time-of-day surge pricing from a major API — 2x listed rates during Beijing peak hours — with V4 GA signaled as imminent

Surge pricing is a structural first: inference is being priced like a capacity-constrained utility, so cost models that assume flat per-token rates need a time-of-day dimension, and latency-tolerant batch workloads should be scheduled into off-peak windows. The alias retirement is a hard migration deadline — audit for pinned deepseek-chat/deepseek-reasoner IDs now.

SOURCE: DEEPSEEK API NEWS

2026-07-15

SIGNAL 04

OpenAI

GPT-Red stays internal: a model trained at flagship post-training scale that OpenAI states will never be released — its output ships only as hardening folded into released models

This is the clearest capability-gating precedent yet: the strongest capabilities may increasingly arrive as vendor-internal advantages you rent the byproducts of, not products you evaluate. Expect competitors to follow, and adjust diligence accordingly — ask vendors what internal-only systems shaped the model you are buying, and demand the quantified safety deltas (0.05% direct-injection failure is now the number to beat).

SOURCE: OPENAI (VENDOR-REPORTED)

2026-07-15

SIGNAL 05

Google

Reported in-place refresh of the Gemma 4 weights, FlashAttention-4 kernels, and chat templates across the HF collection — targeting tool-call JSON consistency, vision OCR, and long-prompt prefill (weakly sourced; single secondary blog)

If confirmed, in-place weight refreshes under an unchanged model name break reproducibility assumptions: self-hosted Gemma 4 operators should re-pull stale local caches and pin content hashes, not model names, in any evaluation or compliance pipeline. Treat this as a caution flag until Google documents it first-party.

SOURCE: EXPLAINX (SECONDARY; NOT YET CONFIRMED FIRST-PARTY)

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 JUL 27

WATCH 01

Kimi K3 weights, license, and tech report

The week's biggest open question: if clean weights and a permissive license land as promised, K3 becomes the strongest open-weights model on record and the open-frontier scorecard flips; if the date slips or the license restricts, the 'open launch without weights' pattern hardens into a playbook.

JUL 24

WATCH 02

DeepSeek V4 GA + hard alias retirement

deepseek-chat and deepseek-reasoner go dark the same window V4 is signaled to arrive ('mid-July' per DeepSeek) — the forced migration will produce the first clean V4 adoption data and test whether surge pricing survives contact with customers.

WEEK OF JUL 20

WATCH 03

GPT-Red technical preprint

Every GPT-Red number is currently vendor-reported; the preprint is the first chance for independent scrutiny of the 84%-vs-13% red-teaming claim and the Fake Chain-of-Thought injection class, which together justify making injection robustness a procurement spec.

BY JUL 31

WATCH 04

Gemini 3.5 Pro slip watch

The leaked Jul 17 target passed — the third slip. Prediction-market odds (~81% by Jul 31 on Polymarket) are the only forward signal available and are speculation-grade; treat any capacity or contract planning around a 2M-context flagship as unanchored until Google commits a date.

NO DATE ANNOUNCED

WATCH 05

Inkling-Small (276B/12B active) weights

The previewed small tier is the practical deployment target for most teams eyeing the Inkling base — its release (and whether the Apache 2.0 terms carry over) determines whether Thinking Machines' open push reaches beyond 600GB-class serving budgets.

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 13

Week 29 of 2026 · July 18, 2026

WEB

brianletort.ai/industry/model

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