THE MODEL PULSE ISSUE 19 · WEEK 35 OF 2026

THE BIG READ

Kodam's practitioner harness result and seat metering, not Hot Chips silicon — the week's model procurement read

KEY SIGNALS THIS PERIOD

5 Models added to the tree

6 Vendors shipped frontier-class

3 Architectural patterns crossed multiple vendors

4 Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

August 29, 2026

Issue 19 · 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.

Vijay Kodam's Aug 24 practitioner bakeoff documents an 8× wall-clock swing on identical Qwen3.8-27B weights from the harness-plus-runtime pairing — Pi+Ollama 34 minutes versus Qwen Code+LM Studio 4h46m, with his own Pi+LM Studio control at 115.8 minutes, so neither the harness nor the inference engine alone explains the spread (grade-3 bakeoff in Agent Techniques proofOfValue[1]); GitHub Copilot promotional credits expire September 1 — see Application Layer pricingShifts for Business and Enterprise credit math. For near-term agent economics, test those practitioner-documented integration and metering effects before treating Hot Chips rack headlines as the binding lever — local wall-clock and datacenter tokens-per-megawatt are different axes, not a ranking. Managed-stack buyers should note OpenAI's GPT-5.6-in-Kiro reports ~82% cost reduction from tier routing alone (software-02).

Five open-weight tree rows landed August 24–28 plus Ox Alpha stealth demand (community-reported, grade 3): GLM-5.3-Flash MIT weights, Qwen3.8-Flash-Next architecture preview, IBM Granite 4.2 at vendor-reported 57.00% SWE-Bench Verified (30B), Tencent Hy4 preview, and Thomson-1.0-Small as a Qwen3.6 derivative. OpenAI integrated GPT-5.6 Sol, Terra, and Luna into AWS Kiro with vendor-reported ~82% cost reduction per successful Terminal-Bench 2.1 task versus an unspecified baseline.

Qwen3.8-Flash-Next ships as an open checkpoint under qwen-community-1.0 while production API SKU Qwen3.8-Flash at $0.16/$0.47 per million tokens with 1M default context is a separate hosted product — procurement must not treat the 360GB artifact and the metered API as one SKU.

IBM Granite 4.2 is the week's clearest enterprise on-prem license play: Apache 2.0, switchable chain-of-thought, native tool calling, and sandbox agentic RL yielding vendor-reported 57.00% SWE-Bench Verified at 30B. Read the terminal-agent row before generalizing that: IBM's own table puts the 30B at 29.24% on Terminal-Bench 2.1, roughly 55 points below GLM-5.3-Flash's vendor-reported 84.3% on the same benchmark, so Granite is a deployability and licensing choice

rather than a capability one. Granite Speech 5.0 Turbo ASR on the same release day reports RTFx (real-time factor — throughput relative to audio duration) near 12,600 on a single H200 per IBM — making speech-to-agent loops viable without a hyperscaler API. Thomson Reuters spent ~$40M continual-learning open weights: the Thomson LLM that reaches customers is a closed deployment built from Alibaba's Qwen3.5-397B open weights, and Thomson-1.0-Small is an open-weight Qwen3.6-35B-A3B derivative — see Application Layer for vertical packaging on Claudeforce and Gemini Enterprise.

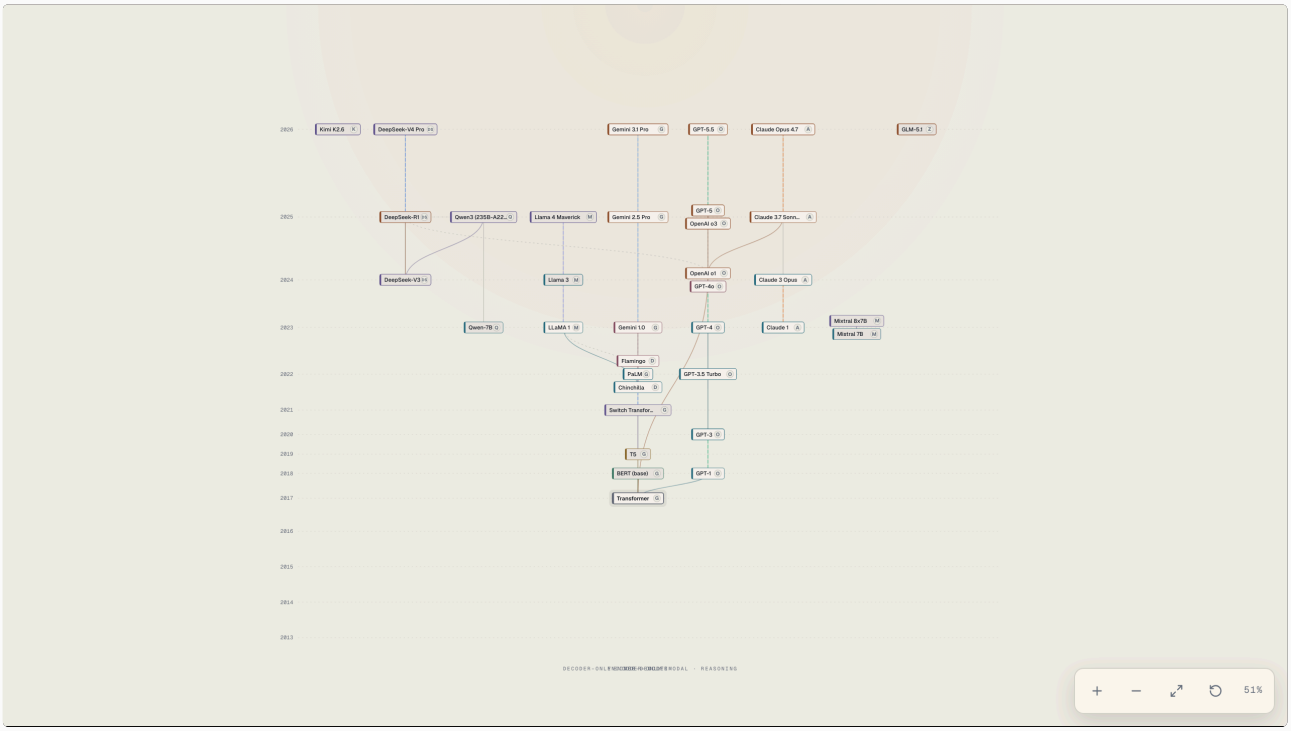
THE BAR

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.

5 model rows added and 1 updated. The canopy widened across open MoE (GLM-5.3-Flash, Hy4 preview), Qwen4 architecture preview (Qwen3.8-Flash-Next), enterprise Apache 2.0 agent stack (Granite 4.2), and legal vertical derivative (Thomson-1.0-Small). GLM-5.3 parent row updated to reflect Flash weight release.



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

ADDED (5)

- glm-5-3-flash
- qwen3-8-flash-next
- granite-4-2-30b
- thomson-1-0-small
- hy4-preview

UPDATED (1)

- glm-5-3

GLM-5.3-Flash confirms Ox Alpha, and Z.ai closed the weights promise days later: the full 753B GLM-5.3 checkpoint published to Hugging Face Aug 27–28 under a bespoke license (Z.AI security review required for Model-as-a-Service operators above $10B revenue), not MIT like Flash. Qwen3.8-Flash-Next is a preview checkpoint — production API is separate SKU Qwen3.8-Flash.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-08-24

RELEASE 01

OpenAI

FRONTIER

REASONING

GPT-5.6 in AWS Kiro

Full GPT-5.6 family (Sol, Terra, Luna) inside AWS Kiro spec-driven coding agent — vendor-reported ~82% lower cost per successful Terminal-Bench 2.1 task versus unspecified baseline

Buyers routing agentic coding through AWS should evaluate Kiro as a bundled harness plus model tier choice, not as raw API pricing alone. OpenAI documents ~82% cost reduction per successful Terminal-Bench 2.1 task in Kiro on its blog; demand baseline definitions before moving production traffic.

SOURCE: OPENAI

2026-08-26

RELEASE 02

Z.AI (Zhipu)

FRONTIER MOE

GLM-5.3-Flash (Ox Alpha reveal)

320B/18B-active natively multimodal MoE MIT weights — six-day stealth run processed 20T+ OpenRouter tokens before identity reveal

Route high-volume coding and agent loops to GLM-5.3-Flash for pilot this quarter if cost dominates, but model post-promo pricing September 9 and independent AA runs after the free stealth period. Serving-hardware claims circulating this week are ungraded: cheap inference under a promotion is not evidence of export-control training independence.

SOURCE: Z.AI

OPEN WEIGHTS

Open-frontier and open-source drops.

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

2026-08-26

RELEASE 01

Z.AI (Zhipu)

OPEN FRONTIER

MOE

GLM-5.3-Flash

MIT-licensed 320B/18B-active multimodal MoE with vendor-reported 84.3% Terminal-Bench 2.1 and AA Intelligence Index 57 at $0.045/task

The artifact is real and downloadable; every headline score is vendor-run or measured during free anonymous access. Teams needing MIT-clean weights for agent pilots should download now; teams needing audited leaderboard position should wait for Artificial Analysis refresh post-promo.

SOURCE: HUGGING FACE

2026-08-26

RELEASE 02

Alibaba

OPEN FRONTIER

MOE

Qwen3.8-Flash-Next

Qwen4 hybrid DeltaNet + Gated Attention preview — 125B/6B active, 512 experts, ~360GB FP checkpoint

Architecture preview for cost-efficient agent routing; self-host only if you can absorb 360GB and qwen-community-1.0 license terms. Most enterprises should start on hosted Qwen3.8-Flash API at $0.16/$0.47 with 1M context rather than this checkpoint.

SOURCE: ALIBABA CLOUD

2026-08-25

RELEASE 03

IBM

OPEN FRONTIER

REASONING

IBM Granite 4.2 (30B)

Apache 2.0 agentic RL family — vendor-reported 57.00% SWE-Bench Verified and 29.24 Terminal-Bench 2.1 at 30B

Default open-weight pick for regulated on-prem agent estates this quarter. Pair with Granite Speech 5.0 Turbo for voice-to-agent loops without frontier API dependency. Require independent SWE-Bench reproduction before substituting for closed frontier on highest-risk code paths.

SOURCE: IBM RESEARCH

2026-08-28

RELEASE 04

Tencent

OPEN FRONTIER

MOE

Tencent Hy4 Preview

770B/49B-active MoE, 1M context, Apache 2.0 — internal blind eval 2.99/4.00 on 203 engineering tasks

Strongest Chinese open MoE drop of the week on license terms, but internal eval is vendor-run. Pilot on FP8 vLLM images if you already run Tencent stack; otherwise wait for independent coding leaderboard entries.

SOURCE: TENCENT

2026-08-24

RELEASE 05

Thomson Reuters

SPECIALIST MOE

Thomson-1.0-Small

Qwen3.6-35B-A3B derivative on Hugging Face under non-commercial academic license — not Apache open weights

Legal teams should treat this as a corpus-tuned derivative for research and non-commercial experimentation, not as a substitute for Thomson's closed-deployment Thomson LLM (itself a continual-learning derivative of Qwen3.5-397B open weights per the Thomson technical report) in production CoCounsel. Editorial forecast: lineage disclosure may become an RFP requirement — the falsifier (two top-tier vertical platforms train fully proprietary frontier models with no open-weight base disclosed, by Q2 2027) is stated once in the Weekly synthesis.

SOURCE: THOMSON REUTERS

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

Open-weight releases bifurcate hosted API SKUs from downloadable checkpoints

`Qwen3.8-Flash vs Flash-Next` `GLM-5.3-Flash stealth vs list API` `Thomson closed vs Small derivative`

Three vendors shipped paired artifacts this week where the production metered endpoint differs from the Hugging Face row — different context defaults, tool interfaces, or licenses. Procurement contracts must name the SKU (hosted API id or checkpoint hash), not the marketing family name, because substitutability failed on every paired release.

SOURCE: ALIBABA CLOUD, Z.AI, THOMSON REUTERS

PATTERN 02

Agentic RL in open-weight post-training is becoming table stakes for coding MoE

`IBM Granite 4.2 sandbox RL` `GLM-5.3-Flash terminal benchmarks` `Hy4 self-optimization`

IBM's Granite 4.2 explicitly trains in sandbox software-engineering, terminal, and web-search environments; Z.ai and Tencent headline terminal and SWE scores on the same week's releases. The architecture shift is not bigger MoE alone — it is RL inside harness-shaped sandboxes. Buyers evaluating open weights for agents should require disclosure of sandbox design, not just parameter counts.

SOURCE: IBM RESEARCH, Z.AI, TENCENT

PATTERN 03

Vertical corpora plus orchestration atop semi-open foundations replaces from-scratch frontier training

Thomson on Qwen3.6 | CoCounsel Claude Agent SDK | Gemini Enterprise verticals | Fuse Terminal GA | Ling 3.0 Flash Fin free window

Thomson Reuters spent ~$40M on continual learning over open weights: Thomson LLM is a closed deployment derived from Qwen3.5-397B and Thomson-1.0-Small an open Qwen3.6-35B-A3B derivative, while CoCounsel retains Claude Agent SDK for agentic workflows (Law.com, Aug 24). Fuse Terminal GA ($999/mo, never-gated API) and Ling 3.0 Flash Fin free through Sep 25 offer finance alternatives to packaged verticals. Regulated vertical moats consolidate on corpus plus platform-governed orchestration — procurement should ask foundation lineage questions in every vertical RFP.

SOURCE: THOMSON REUTERS, LAW.COM

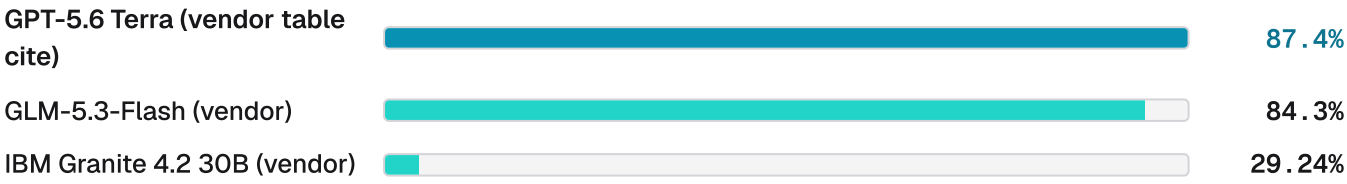
BENCHMARK MOVES

Where the leaderboard moved.

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

TERMINAL-BENCH 2.1 (VENDOR-REPORTED CODING AGENT ROWS) MOVE 01

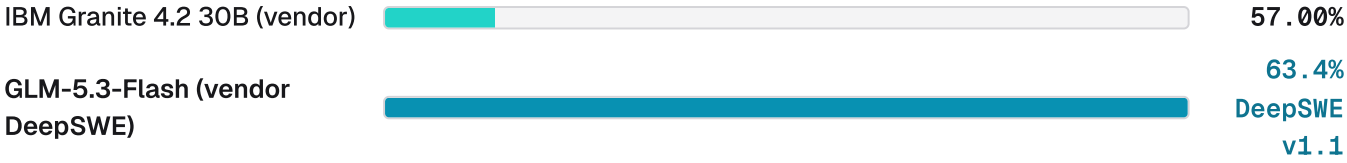
GLM-5.3-Flash publishes vendor-reported 84.3%; IBM Granite 4.2 30B reports vendor-reported 29.24% on the same resolve-rate scale as its 57.00% SWE-Bench Verified row in IBM's own results table — a roughly 55-point gap on the same benchmark, which is the honest read and the reason the open-frontier pick is a licensing decision, not a capability one; GPT-5.6 Terra vendor-reported 87.4% cited on Z.ai comparison table. Harness and tier choice still dominate cross-vendor comparability.



SOURCE: Z.AI, IBM RESEARCH

SWE-BENCH VERIFIED MOVE 02

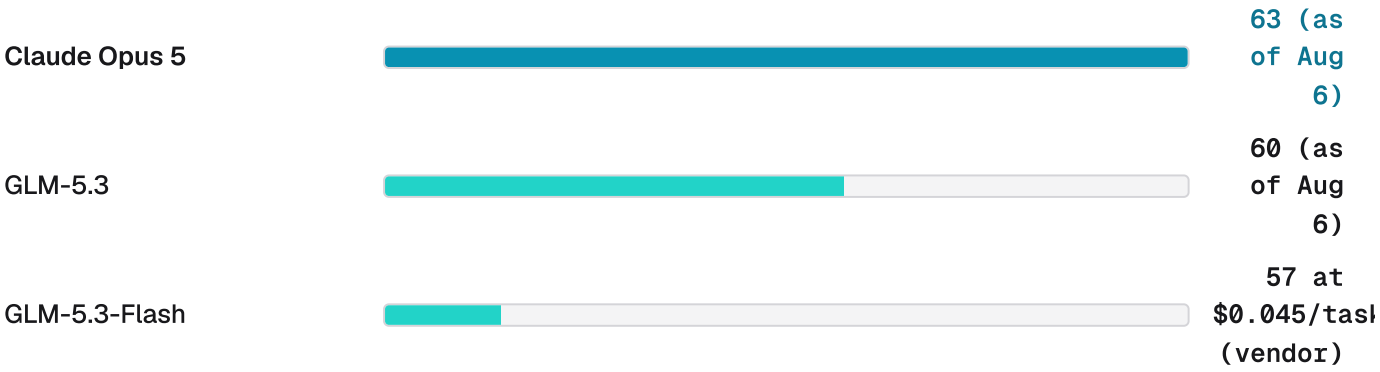
IBM Granite 4.2 30B reports 57.00% — strongest audited-style vendor claim from an Apache 2.0 open release this week; independent reproduction pending.



SOURCE: IBM RESEARCH

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V4.1.1 MOVE 03

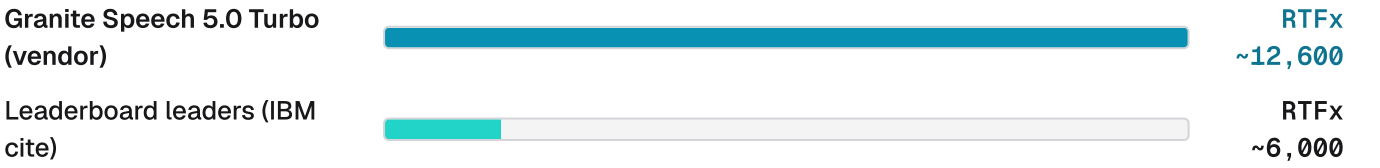
GLM-5.3-Flash scores 57 at vendor-reported $0.045/task during launch window; board unchanged for closed frontier leaders from August 6 refresh.



SOURCE: ARTIFICIAL ANALYSIS, Z.AI

OPEN ASR LEADERBOARD THROUGHPUT (VENDOR-REPORTED) MOVE 04

IBM Granite Speech 5.0 Turbo CTC reports vendor-reported RTFx (real-time factor — audio processed faster than playback) ~12,600 on single H200 — ~2× leaderboard leaders per IBM.



SOURCE: IBM RESEARCH

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Aug 29, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Opus 5**	GPT-5.6 Sol Max	Capability board unchanged from August 6 AA refresh. NVIDIA and OpenAI reframed economics toward agentic tokens-per-megawatt and Kiro tier routing — decide on cost per completed agent task, not list token price alone.
Open frontier	**GLM-5.3-Flash**	IBM Granite 4.2 30B	GLM-5.3-Flash wins on vendor-reported terminal scores and MIT license; Granite wins on Apache 2.0 enterprise deployability and SWE-Bench Verified headline. Neither has independent tracker confirmation at publish — pick license and hosting boundary first, capability second.
Reasoning	**Claude Opus 5**	GPT-5.6 Terra	Terra enters as the efficiency tier inside Kiro with vendor-reported 87.4% Terminal-Bench 2.1 on Z.ai comparison table — route spec-driven coding to Terra when vendor controls harness; Opus remains default for highest-stakes reasoning until independent agentic cost curves land.
Coding	**Claude Fable 5 Max**	GLM-5.3-Flash	GLM-5.3-Flash vendor-reported 84.3% Terminal-Bench 2.1 is not harness-comparable to Fable rows without independent reproduction. Local open-weight coding default shifts to Granite 4.2 30B for Apache estates and GLM-5.3-Flash for MIT high-volume pilots.
Multimodal	**Gemini 3.7 Flash**	GLM-5.3-Flash	GLM-5.3-Flash adds natively multimodal MIT MoE at datacenter scale; Gemini retains hosted enterprise vertical packaging. Qwen3.8-Flash-Next preview adds architecture option for self-host multimodal MoE at extreme footprint.
Edge / small	**Granite Speech 5.0 Turbo**	IBM Granite 4.2 8B	Granite Speech 5.0 Turbo vendor-reported RTFx ~12,600 on H200 is the week's underplayed enterprise angle for voice-to-agent loops. Granite 4.2 3B/8B dense with tool calling

TIER	LEADER	CHALLENGER	READ
			challenges edge leader on enterprise agent features.

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.

2026-08-26

SIGNAL 01

Z.ai

Retired Ox Alpha alias and published GLM-5.3-Flash MIT weights after 20T+ token stealth run; list API $0.15/$0.50 with 50% promo through Sep 9

Stealth free access was a demand test, not steady-state pricing. Renegotiate agent contracts before September 9 promo expiry and require post-promo rate cards in writing.

SOURCE: Z.AI

2026-08-26

SIGNAL 02

Alibaba

Qwen3.8-Flash production API at $0.16/$0.47 per million tokens with 1M default context; Flash-Next open checkpoint separate SKU

Hosted Qwen agent routing at sub-frontier list price with million-token default — pilot BYOK agent loops on Qwen3.8-Flash API before reserving Rubin capacity if harness overhead dominates wall-clock.

SOURCE: ALIBABA CLOUD

2026-08-24

SIGNAL 03

OpenAI

GPT-5.6 family integrated into AWS Kiro with Terra positioned as Terminal-Bench efficiency tier

AWS buyers get vendor-controlled harness plus model tier in one procurement line — shifts evaluation from raw API to integrated agent workflow cost.

SOURCE: OPENAI

2026-08-25

SIGNAL 04

IBM

Granite 4.2 Apache 2.0 with agentic RL; Granite Speech 5.0 Turbo ASR at vendor-reported ~12,600 RTFx

IBM re-enters open-weight agent stack for regulated on-prem — competitive against hosted copilots for enterprises blocked from hyperscaler APIs.

SOURCE: IBM RESEARCH

2026-08-28

SIGNAL 05

GitHub

Promotional Copilot credit pools expire Sep 1 — see Application Layer pricingShifts for Business 3,000→1,900 and Enterprise 7,000→3,900 credit math

Metering contraction on the largest developer surface — teams built on promo allowances face overage unless admins cap spend or route to BYOK open-weight stacks validated this week.

SOURCE: GITHUB CHANGELOG

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.

SEP 3
WATCH 01

FERC (Federal Energy Regulatory Commission) comment deadline on PJM Interim Resource Adequacy Service (ER26-3515)

≥50 MW loads without bring-your-own-capacity may face curtailment-before-residential rules — affects Virginia and PJM hyperscale energization math.

SEP 9
WATCH 02

GLM-5.3-Flash launch promo expires

Steady-state API pricing resets — rerun agent business cases on list $0.15/$0.50 and watch Artificial Analysis provider variance.

SEP 15
WATCH 03

GLM-5.3 license review gate (prediction p90 resolved hit — 753B weights on Hugging Face Aug 27–28)

Full weights shipped under a bespoke non-MIT license — Model-as-a-Service operators above $10B revenue need Z.AI security review before commercial use; watch whether that gate changes provider availability on OpenRouter and Artificial Analysis.

SEP 25
WATCH 04

Ling 3.0 Flash Fin free API window ends

Finance-tuned MoE free hosted access through OpenRouter/Vercel Gateway — last day to benchmark spreadsheet and disclosure-retrieval agents at zero list cost.

OCT 2026
WATCH 05

MetaRoCE OCP specification at Global Summit

Loss-tolerant RDMA transport spec drop — sets whether million-GPU Ethernet stays merchant-open or NVIDIA-codesigned Multiplane wins by default.

Q3 FY2027 PRINT
WATCH 06

NVIDIA Vera Rubin mix disclosure in earnings

CFO guided ~20% datacenter revenue from Rubin in Q3 — separates sell-in allocation from fleet utilization on agentic workloads.

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 19

Week 35 of 2026 · August 29, 2026

WEB

brianletort.ai/industry/model

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