THE MODEL PULSE ISSUE 07 · WEEK 23 OF 2026

THE BIG READ

No new closed frontier shipped; the open/local agent substrate widened underneath it.

KEY SIGNALS THIS PERIOD

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

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

June 6, 2026

Issue 07 · 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.

W23 did not produce the expected Gemini 3.5 Pro GA or a fresh Anthropic/OpenAI frontier release. That absence is the story: Claude Opus 4.8 remains the public closed-frontier leader for coding and agentic work, while Google kept Pro in the June watch window and the model layer's actual shipping activity moved down-stack. JetBrains released Mellum2, an Apache-2.0 12B/2.5B-active MoE designed for low-latency routing, RAG, summarization, validation, and sub-agent calls; NVIDIA released Cosmos 3 as an open physical-AI omni-model with Nano 16B and Super 64B variants; H Company released Holo3.1 with local computer-use sizes and quantized checkpoints. The procurement implication is sharper than another leaderboard reshuffle: production agent systems are becoming portfolios of models. Keep Opus/GPT/Gemini-class models for high-risk reasoning and codebase-scale orchestration, but push cheap, private, repeated sub-agent work into specialized open/local models. The tree delta therefore adds efficient-agent and physical-AI nodes rather than another general chatbot crown.

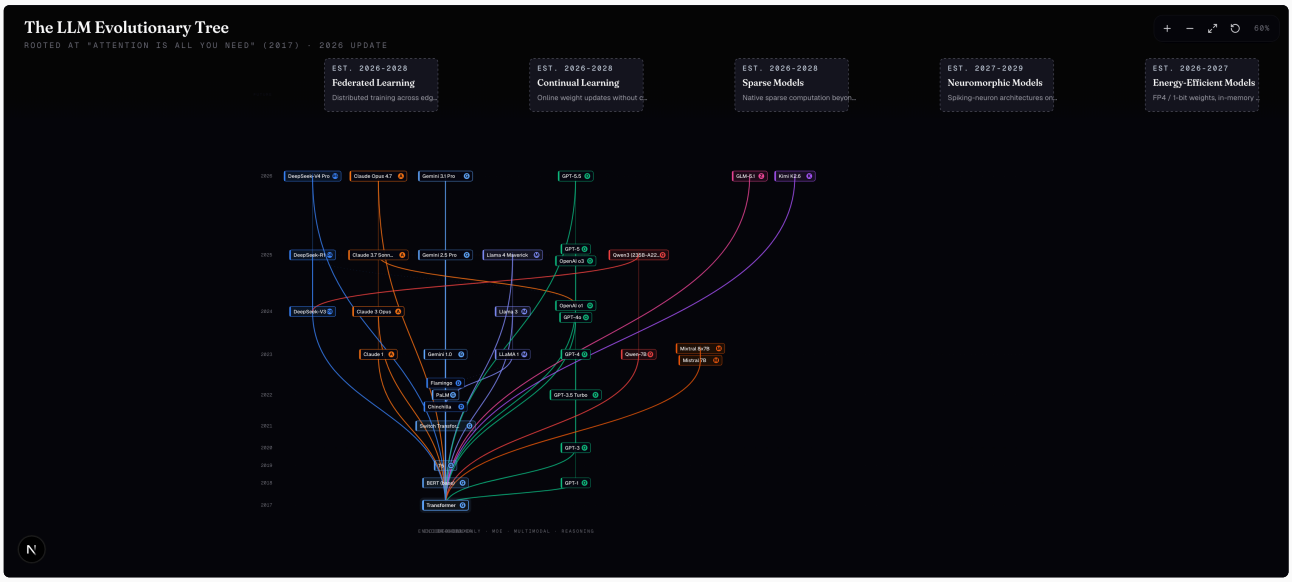
THE BAR

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 W23 additions: Mellum2 for efficient text/code sub-agent workloads, Cosmos 3 for physical-AI omni-modeling, and Holo3.1 for local computer-use agents.



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

ADDED (3)

mellum2
cosmos-3-nano
holo-3-1

UPDATED (0)

No updates to existing rows this period.

Gemini 3.5 Pro remains pending; no tree row is added until Google publishes the GA model card or API identifier.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-06 TARGET

RELEASE 01

Google DeepMind

FRONTIER

REASONING

Gemini 3.5 Pro

Still pending at W23 close: Google has said Pro follows Gemini 3.5 Flash in June, but no public API ID, pricing, or independent benchmark row landed in-window

This is a frontier movement by absence. Buyers waiting for Gemini 3.5 Pro should keep the launch on the June watchlist but should not pause current coding-agent baselines: the public board still has Opus 4.8 leading the closed frontier, and Pro's economics and benchmark profile remain unverified. The likely enterprise split is task routing — Gemini for huge-context/multimodal work, Opus/GPT for coding and agentic reliability — not a universal replacement.

SOURCE: GOOGLE GEMINI 3.5 ANNOUNCEMENT; AI TOOL BOLT JUNE COMPARISON

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.

2026-06-01

RELEASE 01

JetBrains

EDGE / SMALL

MOE

Mellum2

Apache-2.0 12B MoE with 2.5B active parameters per token for low-latency text/code sub-agent workloads

Mellum2 is not trying to win the frontier leaderboard; it is trying to lower the cost of the thousands of routine model calls inside agent systems. Routing, RAG, summarization, validation, and lightweight code tasks are exactly where private deployments want an efficient open model. If the benchmark claims hold, this is a practical procurement node for teams trying to cut orchestration cost without sending every step to a closed flagship.

SOURCE: HUGGING FACE JETBRAINS MELLUM2 LAUNCH

2026-06-01

RELEASE 02

NVIDIA

SPECIALIST

MULTIMODAL

NVIDIA Cosmos 3

Open physical-AI omni-model family combining world generation, physical reasoning, and action generation in Nano 16B and Super 64B variants

Cosmos 3 widens the model tree away from text-only agents and into physical AI. The important shift is a unified model that can reason over world state and action generation rather than stitching separate generation and control pipelines together. Robotics, simulation, and industrial automation teams should evaluate it as synthetic-data and reasoning infrastructure, not as a chatbot substitute.

SOURCE: HUGGING FACE NVIDIA COSMOS 3 LAUNCH

2026-06-02

RELEASE 03

H Company

SPECIALIST

AGENTIC

Holo3.1

Local computer-use agent family with 0.8B, 4B, 9B, and 35B-A3B sizes plus FP8, Q4 GGUF, and NVFP4 checkpoints

Holo3.1 pushes computer-use agents toward deployability: multiple sizes, quantized checkpoints, and local inference targets matter more than a single headline score. Enterprises automating browser, desktop, and internal-tool workflows can now separate privacy-sensitive UI action from hosted frontier reasoning. That supports a two-layer architecture: local CUA for execution, closed frontier for planning and verification.

SOURCE: HUGGING FACE HOLO3.1 LAUNCH

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

Cheap specialist sub-agents

`Mellum2` `Holo3.1-0.8B / 4B / 9B` `Claude Opus 4.8 fast mode`

The agent stack is splitting into high-reasoning planners and cheap repeated workers. Mellum2 and Holo3.1 are purpose-built for the calls that happen hundreds or thousands of times inside a workflow: routing, validation, summarization, UI action, and local execution. Model routers should now budget by step type rather than treating one flagship as the default for every agent call.

SOURCE: HUGGING FACE MELLUM2 AND HOLO3.1 LAUNCHES

PATTERN 02

Physical-AI omni-models

`NVIDIA Cosmos 3 Nano` `NVIDIA Cosmos 3 Super` `Bernini-R renderer`

Open model activity is expanding from language and code into physical-world simulation, video rendering, and action generation. Cosmos 3's combined world generation, physical reasoning, and action generation points toward a branch where synthetic data and robotics workflows become first-class model workloads. That is a different buyer and deployment path than enterprise chat.

SOURCE: HUGGING FACE NVIDIA COSMOS 3 LAUNCH; BYTEDANCE BERNINI-R HUGGING FACE CARD

PATTERN 03

Frontier release gaps measured in weeks

`Claude Opus 4.8` `Gemini 3.5 Pro pending` `GPT-5.6 speculation`

The absence of a new frontier release this week matters because expectations have compressed. Buyers are now tempted to delay procurement for a model that may arrive in days. The practical answer is to separate infrastructure choices from model choice: standardize evaluation harnesses, routers, and cost controls so a June GA can be tested and slotted without freezing current deployments.

SOURCE: GOOGLE GEMINI 3.5 ANNOUNCEMENT; ANTHROPIC OPUS 4.8 ANNOUNCEMENT

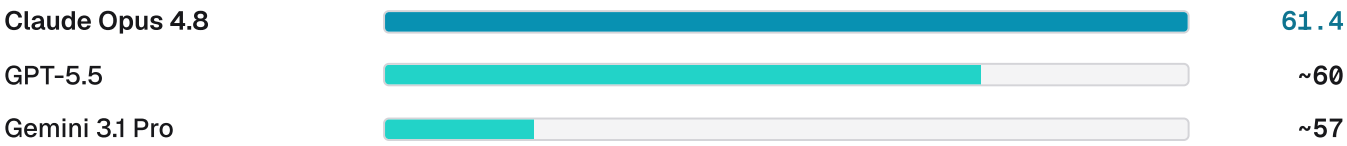
BENCHMARK MOVES

Where the leaderboard moved.

3 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 W23 leaderboard reset; Claude Opus 4.8 remains the public #1 at 61.4 while Gemini 3.5 Pro is still pending GA



SOURCE: ARTIFICIAL ANALYSIS SUMMARIES VIA JUNE MODEL COMPARISONS

SWE-BENCH PRO MOVE 02

Closed frontier still leads coding: Opus 4.8's 69.2% remains the public bar; no new open release in W23 changes the top coding score

SOURCE: AI TOOL BOLT / PRISTREN JUNE BENCHMARK SUMMARIES

LOCAL COMPUTER-USE DEPLOYABILITY MOVE 03

Holo3.1 shifts the measurable axis from one top-line CUA score to model size and quantization availability for local execution

Holo3.1-0.8B	ultra-light local
Holo3.1-9B	balanced local
Holo3.1-35B-A3B	state-of-the-art tier

SOURCE: HUGGING FACE HOLO3.1 LAUNCH

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Jun 6, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	Claude Opus 4.8	GPT-5.5	No W23 reset; Gemini 3.5 Pro remains the watched June challenger rather than a published benchmark row.
Open frontier	DeepSeek V4-Pro	GLM-5.1	No new open frontier text model displaced the April leaders; W23 open activity shifted to specialist/local models.
Reasoning	Claude Opus 4.8	GPT-5.5	Closed reasoning leadership steady while Gemini 3.5 Pro remains pending.
Coding	Claude Opus 4.8	GPT-5.5	Opus 4.8 still owns the visible SWE-Bench Pro lead; Mellum2 matters for cheap sub-agent code/text calls.
Multimodal	Gemini 3.1 Pro	Cosmos 3	Gemini remains the general multimodal reference; Cosmos 3 creates a specialist physical-AI branch.
Edge / small	Mellum2	Holo3.1-9B	Efficient local/sub-agent models were the week's real release activity.

VENDOR SIGNALS

Pricing, gating, deprecation.

3 non-release moves that shift vendor risk — pricing, deprecations, gating decisions, license changes — each with a one-line procurement read.

2026-06

SIGNAL 01

Google DeepMind

Gemini 3.5 Pro remains promised for June with no public API model ID, pricing, or third-party benchmark row by W23 close

Procurement teams should prepare an eval slot but avoid freezing current deployments on an unreleased model. The operating pattern is rapid re-baselining, not launch-date speculation.

SOURCE: GOOGLE GEMINI 3.5 ANNOUNCEMENT AND JUNE DEVELOPER GUIDANCE

2026-06-01

SIGNAL 02

JetBrains

Released Mellum2 under Apache 2.0 with an explicit low-latency production-workload positioning

IDE and enterprise-platform vendors can now point to a plausible open/private model for background text-code tasks. That increases pressure on hosted copilots to justify every closed-frontier call by risk or quality, not habit.

SOURCE: HUGGING FACE JETBRAINS MELLUM2 LAUNCH

2026-06-01

SIGNAL 03

NVIDIA

Released Cosmos 3 on Hugging Face while also announcing Vera Rubin production at GTC Taipei

NVIDIA is binding the model and infrastructure stories together: physical-AI models create demand for the simulation, synthetic-data, and rack-scale compute stack it sells. Buyers should evaluate model capability and deployment substrate together.

SOURCE: HUGGING FACE COSMOS 3 LAUNCH; NVIDIA GTC TAIPEI

WATCHLIST

On the radar next.

3 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.

JUN 7-30

WATCH 01

Gemini 3.5 Pro GA

The first public model card, price row, API ID, and Artificial Analysis pass will determine whether June becomes a true frontier reset or just a routing expansion.

JUN-JUL

WATCH 02

Mythos-class Anthropic availability

Anthropic has publicly framed stronger gated models as pending cyber safeguards. A wider release would change the closed-frontier scorecard more than another Opus point release.

JUN-AUG

WATCH 03

Open/local agent model adoption

Downloads, integrations, and benchmark replications for Mellum2, Cosmos 3, and Holo3.1 will show whether specialist open models are becoming production substrate or just launch-week noise.

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 07

Week 23 of 2026 · June 6, 2026

WEB

brianletort.ai/industry/model

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