THE MODEL PULSE
ISSUE 04 · WEEK 20 OF 2026

THE BIG READ

Frontier text took a breath; the specialist canopy widened — video, on-device multimodal, open world-models — and agent-runtime monetization became the story.

KEY SIGNALS THIS PERIOD

3
Models added to the tree

4
Vendors shipped frontier-class

4
Architectural patterns crossed multiple vendors

4
Benchmark moves

AUTHOR
Brian Letort
BrianLetort.AI

PUBLISHED
May 17, 2026
Issue 04 · 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.

Frontier text models took a breath this week — no new GPT-class or Claude refresh between W19 and Google I/O (May 19-20), and LMArena top-of-board moved less than 1 Elo. But the specialist canopy widened sharply, with three model rows joining the tree across video / embodied, on-device multimodal, and open-source world modeling.

Perceptron Mk1 collapsed video and embodied reasoning costs by 80-90% on May 12, pricing frontier-grade physical-AI perception under Gemini Flash Lite ($0.15 / $1.50 per Mtok, 85.1 EmbSpatialBench, 72.4 RefSpatialBench) and forcing a re-cost of any 2H roadmap that ships robotics, surveillance, or screen-watching agents. NVIDIA's open-weights SANA-WM (May 15, Apache 2.0) put a minute-scale 720p world model on a single RTX 5090 in 34 seconds — robotics, AV simulation, and synthetic-data teams can now in-house what they previously rented from closed video APIs. OpenBMB's MiniCPM-V 4.6 1.3B (May 11) reset the low end of multimodal SLMs at 262K context with native iOS / Android / HarmonyOS deployment, compressing on-device feature timelines from quarters to weeks.

On the platform layer, the unit of competition shifted from model intelligence to agent-runtime monetization. Anthropic's Code with Claude conference (May 6-12) made Managed Agents — Dreaming (between-session memory consolidation), Multiagent Orchestration, Outcomes (rubric-graded), and signed Webhooks — a hosted service; Claude Platform on AWS hit GA May 11; xAI shipped Grok Build CLI on May 14 with a SuperGrok Heavy tier at $300/mo; OpenAI's realtime voice trio (GPT-Realtime-2 / Translate / Whisper, added W19) went broadly available May 11 with 128K context and adjustable reasoning effort. Architects building bespoke agent stacks should justify build vs buy explicitly in 2H plans.

The W18 Pentagon vendor-disqualification thread compounded. UK AISI published paired evaluations May 13 confirming autonomous offensive-cyber capability is now doubling every 4.7 months (aligned with METR's 4.2-month SWE figure), and Anthropic emailed Max-20x subscribers a June 15 policy that moves Claude Code

third-party agents off subscription rate limits — 12x-175x effective price increase per workload. Capability gating, federal procurement, and pricing structure are now durable vendor-risk dimensions; bench scores alone no longer settle a procurement decision.

For the May 19-26 window, the load-bearing catalysts are Google I/O 2026 (Gemini 3.2 Flash / Pro confirmation, fabric session), NVIDIA Q1 FY27 (May 20), Microsoft Build (May 19-22), the Samsung HBM4 walkout starting May 21, and any Anthropic / OpenAI counter-launch following I/O.

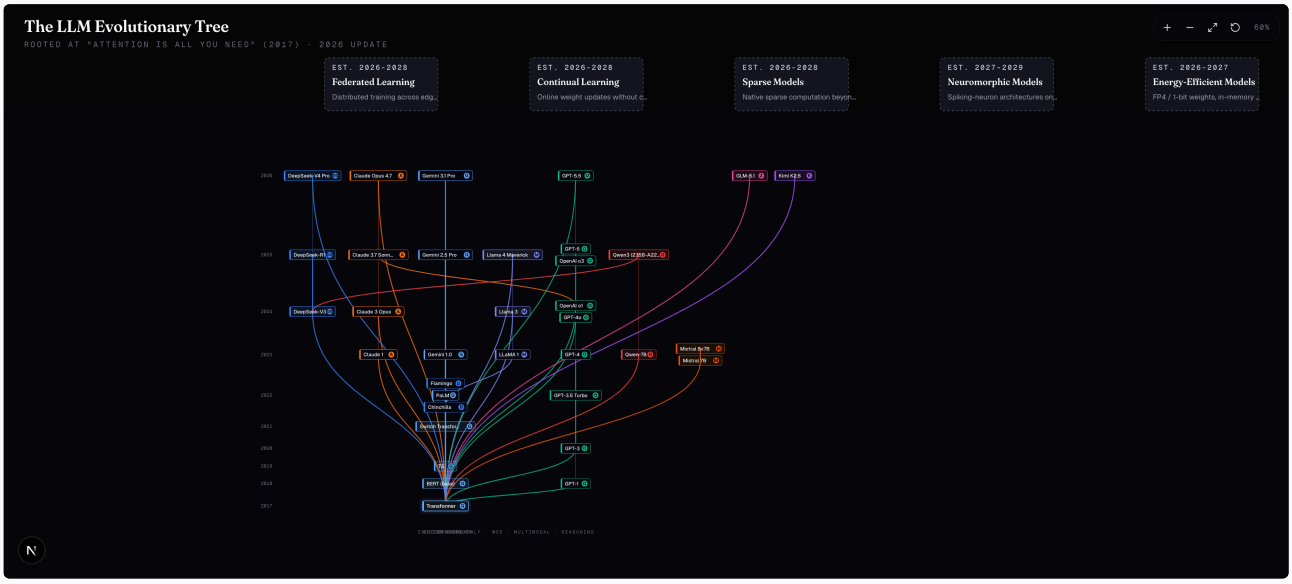
THE BAR

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.

3 model rows added across 3 vendors in a single 7-day window; the canopy widened in specialist multimodal (video / embodied), open-source world modeling, and on-device multimodal SLMs. Closed-source text frontier was quiet ahead of Google I/O.



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

ADDED (3)	UPDATED (3)
minicpm-v-4-6	gpt-realtime-2
perceptron-mk1	gpt-realtime-translate
sana-wm	gpt-realtime-whisper

OpenAI realtime voice family (added W19) hit broad public availability May 11 — flipped status to GA in updated rather than re-adding. Frontier text held flat: GPT-5.5 (xhigh) at AA Index 60 unchanged; Claude Opus 4.7 (Adaptive Max) at 57; LMArena top-of-board moved less than 1 Elo.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-05-12

RELEASE 01

Perceptron

SPECIALIST

MULTIMODAL

Perceptron Mk1

Re-cost any 2H roadmap that ships video, robotics, or screen-watching agents — frontier-grade perception just landed below Gemini Flash Lite.

Perceptron Mk1 matches Gemini Pro / Claude / GPT-class on video and embodied reasoning at $0.15 / $1.50 per Mtok — 80-90% under incumbent multimodal pricing — and scores 85.1 EmbSpatialBench / 72.4 RefSpatialBench. Architects building real-time physical-AI pipelines (manufacturing QA, surveillance, robotics policy verification) should pilot it this quarter; procurement teams should treat it as a credible second source against Google for high-volume video workloads.

SOURCE: PERCEPTRON.INC/BLOG/INTRODUCING-PERCEPTRON-MK1, VENTUREBEAT, BUSINESSWIRE

2026-05-11

RELEASE 02

OpenAI

FRONTIER

MULTIMODAL

GPT-Realtime-2 / Translate / Whisper (suite GA)

Voice agents now get GPT-5-class reasoning at production scale — refresh any call-center or IVR roadmap that still assumes pre-reasoning realtime APIs.

OpenAI's realtime trio shifted from W19 announcement to broad API availability on May 11 with 128K context (up from 32K), adjustable reasoning effort, and parallel tool calls with audible status. Procurement should re-RFP call-center and accessibility workflows that were locked in at last-gen realtime pricing; competitors (Anthropic, Google, Inworld) will have to answer on reasoning-in-voice within 90 days.

SOURCE: OPENAI.COM/INDEX/ADVANCING-VOICE-INTELLIGENCE-WITH-NEW-MODELS-IN-THE-API

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.

2026-05-11

RELEASE 01

OpenBMB

EDGE / SMALL

MULTIMODAL

MiniCPM-V 4.6 1.3B

On-device multimodal just got serious — push a tier of vision / OCR / screen-understanding workloads off cloud APIs before Q3 budget locks.

1.3B params, 262K context, runs natively on iOS / Android / HarmonyOS with quantized variants (GGUF, BNB, AWQ, GPTQ) shipped day one. Hits AA Intelligence Index 13 — beats Qwen3.5-0.8B at 19x lower token cost and matches Qwen3.5 2B on many vision tasks. Operators should pilot offline vision agents (warehouse, retail, field service) and reconsider any cloud-locked on-screen-understanding contract; investors should mark another point on the SLM / edge cost curve.

SOURCE: ARTIFICIALANALYSIS.AI/ARTICLES/OPENBMB-LAUNCHES-MINICPM-V-4-6-1-3B-INSTRUCT, HUGGINGFACE.CO/OPENBMB/MINICPM-V-4.6

2026-05-15

RELEASE 02

NVIDIA

SPECIALIST

MULTIMODAL

SANA-WM

Open-weight world models now fit on one GPU — robotics, simulation, and synthetic-data teams can in-house what they previously rented from closed video APIs.

2.6B Hybrid Linear Diffusion Transformer that produces 60-second 720p clips with 6-DoF camera control in 34 seconds on a single RTX 5090 (NVFP4), a 36x throughput gain over prior open baselines. Architects building robotics pretraining, AV simulation, or content-pipeline tooling should fork SANA-WM rather than sign multi-year deals with closed video APIs; procurement should ask their current video-gen vendor what justifies a 36x throughput premium.

SOURCE: ARXIV.ORG/ABS/2605.15178, NVLABS/SANA GITHUB

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

Specialist video / embodied tier breaks out at frontier quality, sub-Flash-Lite cost

`Perceptron Mk1 (closed, May 12)` `MolmoAct 2 (Ai2, May 5 – adjacent)`

Two physical-AI releases inside the 14-day grace window (Perceptron Mk1 closed, MolmoAct 2 open) signal that video / embodied reasoning is splitting off from general multimodal frontier and pricing under it. Procurement teams that lumped these workloads into a Gemini or Claude SKU should re-segment them in 2H planning; architects should expect a two-tier multimodal stack (general LMM + specialist physical AI) by year-end.

SOURCE: VENTUREBEAT.COM/TECHNOLOGY/PERCEPTRON-MK1-SHOCKS-WITH-HIGHLY-PERFORMANT-VIDEO-ANALYSIS-AI-MODEL

PATTERN 02

Hybrid linear attention compresses long-context cost without quality cliff

`SANA-WM (Hybrid Linear Diffusion Transformer)`

`Subquadratic SubQ (May 5, 12M context, verification pending)`

NVIDIA's SANA-WM uses frame-wise Gated DeltaNet plus softmax attention to deliver minute-scale video on one GPU; Subquadratic's SubQ pushes the same idea to 12M-token text context with claimed ~1000x compute reduction. Operators with long-document, long-video, or long-session agentic workloads should add at least one subquadratic-attention model to their 2H pilot list — the cost curve for ultra-long context is moving faster than transformer-only roadmaps assume.

SOURCE: ARXIV.ORG/ABS/2605.15178; SUBQUADRATIC.AI

PATTERN 03

Edge multimodal SLMs ship native quantization + OS coverage out of the box

`MiniCPM-V 4.6 1.3B`

OpenBMB shipped GGUF / BNB / AWQ / GPTQ quantizations and reference deployments for iOS, Android, and HarmonyOS on day one — the friction between 'open-weights release' and 'phone build' has collapsed. Operators planning on-device features can compress timelines from quarters to weeks; vendors who still ship weights without quantization or mobile runtime support look behind the curve.

SOURCE: HUGGINGFACE.CO/OPENBMB/MINICPM-V-4.6

PATTERN 04

Managed agent runtimes (memory, orchestration, outcome graders) move from preview to hosted product

`Anthropic Claude Managed Agents — Dreaming, Multiagent Orchestration, Outcomes, Webhooks`

Anthropic's Code with Claude (May 6-12) made Dreaming (between-session memory curation), coordinator-subagent orchestration, and rubric-graded Outcomes a hosted service — and Claude Platform on AWS hit GA May 11. Architects building agent stacks should stop rolling these primitives in-house unless they have a clear differentiator; procurement should ask their LLM vendor for an explicit answer on agent-runtime feature parity.

SOURCE: CLAUDE.COM/BLOG/CODE-W-CLAUDE-SF-2026-SF

BENCHMARK MOVES

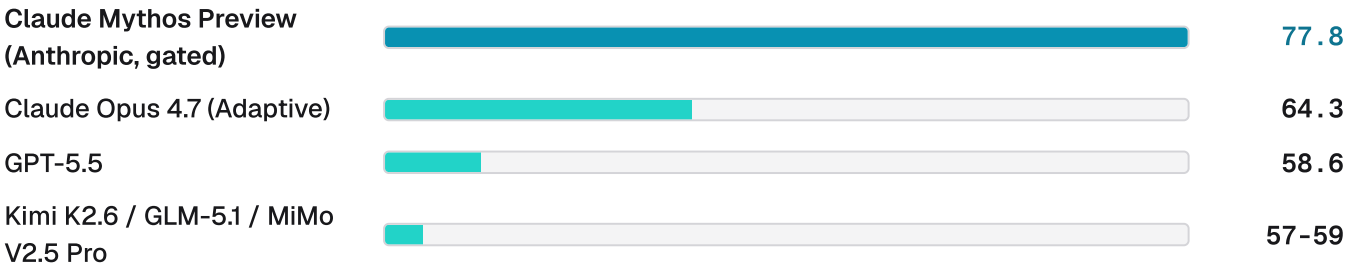
Where the leaderboard moved.

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

SWE-BENCH PRO (AGENTIC CODING)

MOVE 01

Claude Mythos Preview extends its lead with Anthropic occupying the top two slots; GPT-5.5 trails by ~19 points — re-anchor coding-agent procurement on Claude unless GPT-5.5 Instant variant changes pricing.

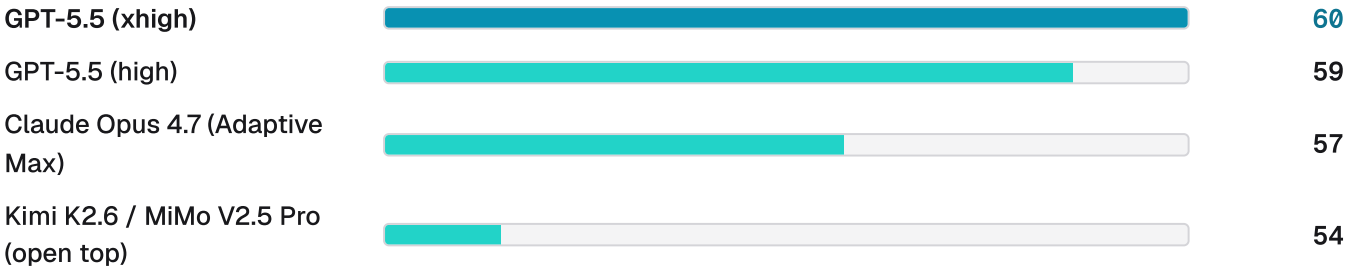


SOURCE: BENCHLM.AI/BENCHMARKS/SWEPRO

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V4 (COMPOSITE)

MOVE 02

GPT-5.5 xhigh leads at 60; GPT-5.5 high 59; Claude Opus 4.7 (Adaptive Max) 57 — Anthropic and OpenAI within a 3-point band, no clear frontier text leader heading into Google I/O.

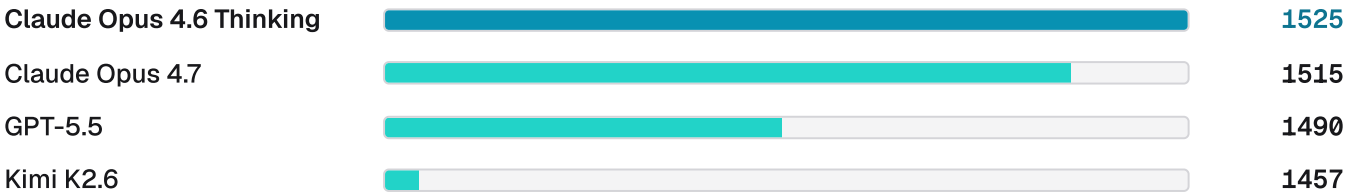


SOURCE: ARTIFICIALANALYSIS.AI/EVALUATIONS/ARTIFICIAL-ANALYSIS-INTELLIGENCE-INDEX

LMARENA OVERALL TEXT (MAY-16 SNAPSHOT)

MOVE 03

Anthropic occupies top 5 (Opus 4.6 Thinking, Opus 4.6, Opus 4.7, Opus 4.7 Thinking, Sonnet 4.6); GPT-5.5 at #6 / #7 — text leaderboard moved less than 1 Elo this week, confirming a quiet text-frontier holding pattern.

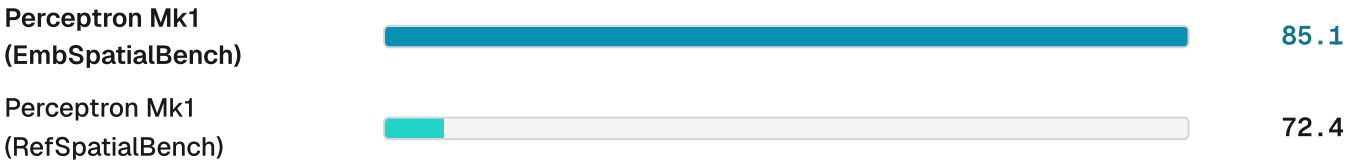


SOURCE: HUGGINGFACE.CO/DATASETS/LMARENA-AI/LEADERBOARD-DATASET

EMBSPATIALBENCH / REFSPATIALBENCH (EMBODIED SPATIAL REASONING)

MOVE 04

Perceptron Mk1 sets a new ceiling for affordable spatial reasoning — leaders on closed multimodal bench at a fraction of incumbent pricing, opening physical-AI workloads that previously needed Gemini Pro / Claude Opus.



SOURCE: VENTUREBEAT.COM/TECHNOLOGY/PERCEPTRON-MK1-SHOCKS-WITH-HIGHLY-PERFORMANT-VIDEO-ANALYSIS-AI-MODEL

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of May 17, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**GPT-5.5 (xhigh) — AA Index 60**	Claude Opus 4.7 (Adaptive Max) — AA Index 57; LMArena leader	Three-point band at the top — no clear winner ahead of Google I/O; default by use case, not vendor.
Open frontier	**GLM-5 (Z.ai, 744B / 40B active) — top open-source on Vending Bench 2**	DeepSeek V4 Pro (1.6T / 49B active) — V4.1 with full-modal coverage scheduled June	Open frontier flat this week; DeepSeek V4.1 in June is the next catalyst for the open-vs-closed lever.
Reasoning	**Claude Opus 4.7 — leads GDPval-AA at 1753 Elo**	Kimi K2.6 — top open-weight reasoning at 1457 LMArena Elo	Anthropic still owns enterprise reasoning by ~80+ Elo; Moonshot leads the open side.
Coding	**Claude Mythos Preview — 77.8 SWE-Bench Pro (BenchLM)**	Claude Opus 4.7 (Adaptive) — 64.3 SWE-Bench Pro; GPT-5.5 at 58.6	Anthropic owns the top two coding slots; Mistral Medium 3.5 (77.6 SWE-Verified) leads the open challenger track.
Multimodal	**Gemini 3.1 Pro — 57.2 AA Index, top vision arena**	Perceptron Mk1 (video / embodied specialist) — Gemini-Pro-class at 80-90% lower cost	General LMM leader unchanged; specialist multimodal split is forming under it (video / embodied + on-device).
Edge / small	**Gemma 4 E4B / 31B Dense — HF top trending, #3 LMArena open-leader**	MiniCPM-V 4.6 1.3B — on-device multimodal with 262K context, native iOS / Android	Edge tier added a credible multimodal SLM this week; sub-2B with 200K+ context is the new normal.

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.

2026-05-13
SIGNAL 01
Anthropic

Max-20x policy change effective June 15: Claude Agent SDK, `claude -p`, GitHub Actions, third-party agents move to separate $20-$200/mo metered credit at API list prices — 12x-175x effective price increase per workload

Ends Claude Code subscription arbitrage. Architects running Claude Code at scale must model true API-rate burn before the June 15 cutover or shift workloads to alternative coding agents (OpenAI Codex, xAI Grok Build, Cursor). Procurement should reopen agent-runtime contracts assuming Anthropic's monetization stance is structural; the third-party agent ecosystem (OpenClaw, T3 Code, Conductor, Zed, Jean) faces a churn test.

SOURCE: SUPPORT.CLAUDE.COM, VENTUREBEAT, XDA-DEVELOPERS

2026-05-11
SIGNAL 02
Anthropic

Claude Platform on AWS hits GA

Enterprise procurement teams already standardized on AWS can now buy Claude (including Managed Agents) through AWS billing, IAM, and authentication — removes the last friction excuse not to deploy Claude alongside or instead of Bedrock-native models; expect Anthropic enterprise pipeline to compress this quarter.

SOURCE: CLAUDE.COM/BLOG/CLAUDE-PLATFORM-ON-AWS

2026-05-06 + 2026-05-12
SIGNAL 03
Anthropic

Code with Claude conference — Managed Agents (Dreaming, Multiagent Orchestration, Outcomes, Webhooks) shipped as hosted service

Anthropic now ships an opinionated hosted agent runtime including memory consolidation, coordinator-subagent orchestration, rubric-graded outcomes, and HTTPS-signed webhooks — competing directly with in-house and open agent frameworks. Architects building bespoke agent stacks should justify build vs buy explicitly in 2H plans.

SOURCE: CLAUDE.COM/BLOG/CODE-W-CLAUDE-SF-2026-SF

2026-05-11

SIGNAL 04

Alibaba

Qwen agentic shopping live across all 4B+ Taobao products

Largest production deployment of an LLM-driven conversational commerce surface to date — sets a real-world pattern for embedded agentic UX and gives Alibaba a data moat on conversational purchase signal ahead of the 618 festival. Western retail platforms should benchmark and respond within two quarters.

SOURCE: ALIBABACLOUD.COM/BLOG/ALIBABA-OPENS-ALL-OF-TAOBAO-TO-QWEN-AI-USHERING-IN-A-NEW-AGENTIC-SHOPPING-EXPERIENCE

2026-05-14

SIGNAL 05

xAI

Grok Build coding-agent CLI launched, SuperGrok Heavy ($300/mo) tier

xAI enters the coding-agent CLI race (vs Claude Code, Codex, Cursor) with a local-first architecture and plan / review / approve model — the IDE / agent surface is now a competitive battleground; engineering managers should reassess developer tooling RFPs rather than assume current vendor lock-in holds through year-end.

SOURCE: CIODIVE.COM/NEWS/XAI-CODING-AGENTS-GROK-BUILD

2026-05-09

SIGNAL 06

DeepSeek

DeepSeek V4 image-recognition feature live; V4.1 scheduled June with full-modal coverage + MCP

DeepSeek closes the multimodal capability gap incrementally rather than waiting for a V4.1 step-change — and the June V4.1 commitment includes MCP support and enterprise toolchain, signaling a clear pivot from research demo to enterprise procurement target. Vendor managers should add DeepSeek to RFP shortlists rather than treating it as an experimental shadow.

SOURCE: CNTECHPOST.COM/2026/05/09/DEEPSEEK-ROLL-OUT-IMAGE-RECOGNITION-FEATURE-AHEAD-V4-1-UPDATE

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.

MAY 19-20
WATCH 01

Google I/O 2026 — Gemini 3.2 Flash expected launch (already leaked)

Gemini 3.2 Flash surfaced in iOS app and Eleuther Arena May 5 with ~92% of GPT-5.5 capability at ~$0.25 / $2.00 per Mtok; if confirmed at I/O, becomes the new price anchor for the cheap-frontier tier and forces OpenAI / Anthropic price moves within 60 days. Watch the fabric session for any second-hyperscaler MRC adoption.

MAY 19-26
WATCH 02

Anthropic / OpenAI counter-launch window post-I/O

Both Anthropic (Mythos full launch) and OpenAI (any 5.5.x refresh) historically respond to Google I/O within 7-10 days — high probability of a frontier-text move that will reset the LMArena top of the table; procurement contracts signed this week should leave headroom.

MAY 19 - JUNE 9
WATCH 03

DeepSeek V4.1 — multimodal + MCP + enterprise toolchain

DeepSeek committed publicly to a June V4.1 release with full-modal (image, audio) coverage and Model Context Protocol support — this is the model that decides whether DeepSeek graduates to enterprise RFP shortlists or stays a research curiosity.

MAY 19 - JUNE 16
WATCH 04

Kimi K3 (Moonshot) — 1M context, Kimi Linear attention

Prediction markets show ~74% probability of release inside ~6 weeks; rumored 3-4T parameters and Kimi Linear attention would put the next open frontier reasoning ceiling notably above current open SOTA — open-weights buyers should hold spend decisions if possible.

MAY 19 - JUNE 30
WATCH 05

Meta MSL — Llama 5 or Muse Spark v2

Muse Spark (April 8) was MSL's first model and replaced the Llama frontier line; Meta has yet to ship a second MSL model or clarify Llama 5 status. Any move from Meta resets the open vs closed Meta strategy debate and the Llama ecosystem's roadmap; Llama-derivative shops should plan two scenarios.

MAY 19 - JUNE 30

WATCH 06

Frontier voice response from Anthropic / Google

OpenAI's realtime trio with GPT-5-class reasoning (broadly available May 11) puts pressure on Anthropic and Google to ship a comparable voice-with-reasoning API or cede the call-center / accessibility / multilingual real-time segment for two quarters.

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 04

Week 20 of 2026 · May 17, 2026

WEB

brianletort.ai/industry/model

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