THE MODEL PULSE ISSUE 05 · WEEK 21 OF 2026

THE BIG READ

Six new tree rows. The loudest model-layer week of Q2 — Gemini 3.5 + Omni, Qwen3.7-Max in the global top 5, three open-weights drops that reset what 'open' means.

KEY SIGNALS THIS PERIOD

6	5	5	5
Models added to the tree	Vendors shipped frontier-class	Architectural patterns crossed multiple vendors	Benchmark moves

AUTHOR
Brian Letort
BrianLetort.AI

PUBLISHED
May 23, 2026
Issue 05 · 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.

W21 (May 18-23, 2026) was the loudest model-layer week of Q2. Google's I/O turned out larger than expected — the long-anticipated 'Gemini 3.2 Flash' shipped as Gemini 3.5 Flash, paired with native multimodal video generation in Omni Flash and the Antigravity 2.0 standalone agent IDE; Alibaba's same-week Cloud Summit pushed Qwen3.7-Max into the global AA Intelligence Index top 5 (a first for a Chinese model at 56.6, ahead of Gemini 3.5 Flash); three open-weights drops — Cohere Command A+ (first Apache 2.0 frontier-adjacent MoE), Microsoft Fara1.5 (browser agents that out-score OpenAI Operator and Gemini Computer Use), Tencent Hy-MT2 (translation MoE that runs offline on a phone) — reset what 'open' means at the frontier-adjacent tier.

The structural shift underneath all six launches is the same: every flagship was engineered for sustained agentic loops first and chat second, with 1M-token context now table stakes and speed-per-task on the Pareto frontier replacing peak-intelligence as the procurement metric. Gemini 3.5 Flash hits Terminal-Bench 76.2% / MCP Atlas 83.6% at ~280 output tokens per second — outperforming prior-gen Gemini 3.1 Pro on agentic benchmarks at 4x throughput. Qwen3.7-Max demonstrated a 35-hour autonomous run with 1,158 tool calls. The unit of competition shifted from raw intelligence to runtime monetization and ecosystem lock-in.

NVIDIA's Q1 FY27 print (May 20: $81.6B revenue, Vera Rubin announced for Q3 ship, Vera CPU already in OpenAI / Anthropic / Oracle / xAI hands, $145B supply commitments) confirmed the compute supply chain that makes all this possible is still ramping, not peaking. Huang named Anthropic alongside the hyperscalers as a Blackwell deployment customer — the compute-side demand matches the equity bid behind Anthropic's $30B / $900B round closing imminently per Bloomberg (May 22).

The specialist canopy widened structurally. Cohere Command A+ (218B / 25B active MoE, Apache 2.0, runs on 2x H100s) closed the open-vs-closed gap on enterprise RAG and tool-use with τ^2-Bench Telecom jumping 37% to 85%

generation-over-generation; AA-Omniscience Non-Hallucination #1 at 86%. Microsoft Fara1.5-27B (open-weight, Qwen3.5 fine-tune) scored 72% Online-Mind2Web — beating OpenAI Operator (58.3%) and Gemini Computer Use (57.3%) — collapsing the per-task cost for browser-agent fleets. Tencent Hy-MT2 at 30B-A3B runs offline on a phone after 1.25-bit quantization.

Notable absences: no DeepSeek R3, no Kimi K3, no Llama 5, no Grok 4.4 / 4.5, and Anthropic's I/O-week response was infrastructure (self-hosted Managed Agent sandboxes, MCP tunnels) rather than a new model. For the May 25 - June 13 window, the load-bearing catalysts are Anthropic's round close, the Samsung-union ratification vote May 27-28, Microsoft Build (June 2-3), Computex Taipei + NVIDIA GTC Taipei (June 1-5), and any frontier-text counter-launch from Anthropic / OpenAI.

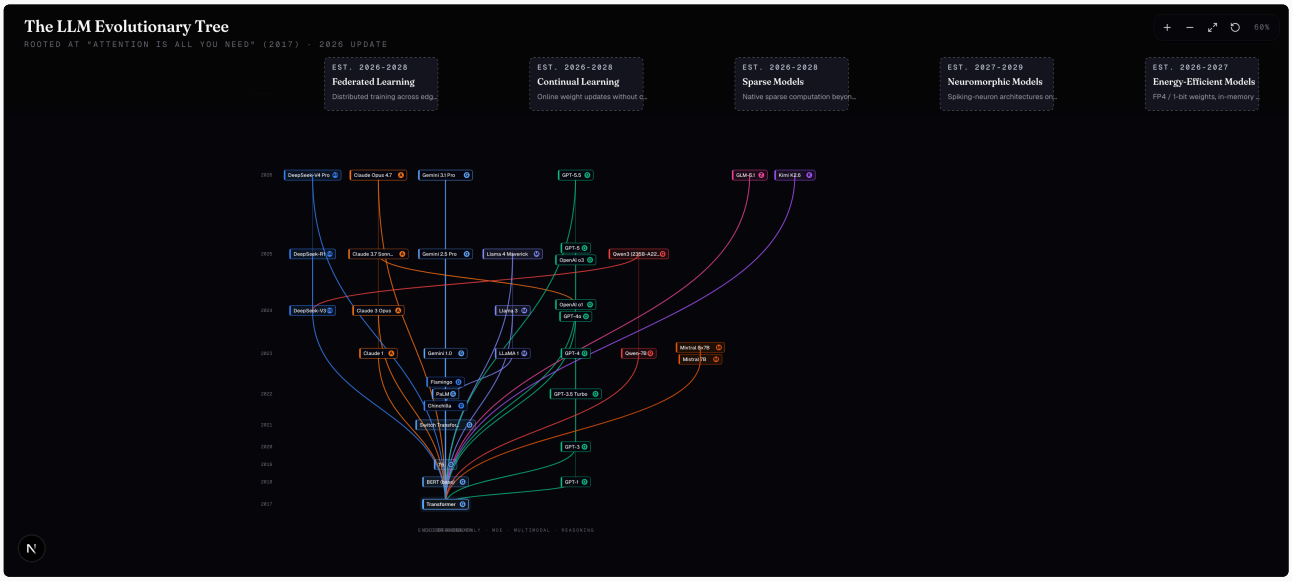
THE BAR

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.

6 model rows added this week across 6 vendors. Canopy widened across Google's I/O drop (Gemini 3.5 Flash, Omni Flash), Alibaba's Cloud Summit response (Qwen3.7-Max — first Chinese model in AA Index top 5), Cohere's first Apache 2.0 frontier-adjacent MoE (Command A+), Microsoft browser-agent open weights (Fara1.5), and Tencent translation MoE (Hy-MT2). Speed-per-task replaces peak-intelligence as the dominant procurement metric.



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

ADDED (6)

- gemini-3-5-flash
- gemini-omni-flash
- qwen-3-7-max
- cohere-command-a-plus
- fara-1-5-27b
- hy-mt2-30b-a3b

UPDATED (0)

No updates to existing rows this period.

I/O + Alibaba Cloud Summit + NVIDIA Q1 FY27 + Cohere / Microsoft / Tencent open-weights all stacked into five business days. This was the densest model-layer week of the quarter. Microsoft Build (June 2-3) sits outside the window, so Microsoft's MAI silence on the platform layer is not yet a signal.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-05-19

RELEASE 01

Google DeepMind

FRONTIER

MULTIMODAL

Gemini 3.5 Flash

Buyers should rerun agentic-coding RFPs — a Flash-tier model now out-codes prior-gen Pro on most agent benchmarks at 4x speed.

Gemini 3.5 Flash hits Terminal-Bench 2.1 76.2%, MCP Atlas 83.6%, and OSWorld 78.4% — beating Gemini 3.1 Pro on almost every agentic eval while running at ~280 output tokens per second, the new Pareto frontier on speed-vs-intelligence. Architects optimizing for sustained agent loops (not chat) should re-baseline cost-per-task within 30 days; Flash pricing is up 3x ($1.50/$9 per Mtok) but per-task cost is often half of frontier peers because of latency and token economy.

SOURCE: DEEPMIND.GOOGLE/MODELS/GEMINI/FLASH, BLOG.GOOGLE

2026-05-19

RELEASE 02

Google DeepMind

FRONTIER

MULTIMODAL

Gemini Omni Flash

Marketing / ad / content-ops leaders should pilot an Omni workflow this quarter — natively multimodal video gen is now in-product, not a research demo.

Omni Flash ingests text / image / audio / video and outputs grounded 10-second video clips with conversational editing — landing simultaneously in the Gemini app, Google Flow, YouTube Shorts, and YouTube Create. The 'create anything from any input' framing is real because the model carries Gemini's world-knowledge through the generation, unlike Veo's text-to-video pipeline. API access lands in weeks; enterprises that ignore this until then will be six weeks behind on creative-ops tooling.

SOURCE: BLOG.GOOGLE/INNOVATION-AND-AI/MODELS-AND-RESEARCH/GEMINI-MODELS/GEMINI-OMNI

2026-05-20

RELEASE 03

Alibaba

FRONTIER

REASONING

Qwen3.7-Max

China-sovereignty buyers have a top-5 global intelligence option in-country — add Qwen3.7-Max to closed-model bake-offs this quarter.

First Chinese model in the Artificial Analysis Intelligence Index top 5 (56.6, ahead of Gemini 3.5 Flash's 55.3); 1M context, 35-hour autonomous tool-use run demonstrated on a real engineering task. Closed-weight strategy and Alibaba Cloud-only access (with homegrown Zhenwu M890 silicon underneath) explicitly target enterprises that want frontier intelligence without US-vendor dependence. The 3-4 point gap to GPT-5.5 has closed enough that 'use Western frontier' is now a debatable position, not a default.

SOURCE: ALIBABACLOUD.COM/BLOG, MARKTECHPOST.COM, ARTIFICIAL ANALYSIS

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.

2026-05-20

RELEASE 01

Cohere

OPEN FRONTIER

MOE

Command A+

Sovereign-AI and data-residency buyers should add Command A+ to their on-prem shortlist — Cohere just shipped the first Apache 2.0 frontier-adjacent MoE that fits on two H100s.

218B total / 25B active MoE, multimodal with citation grounding, W4A4 quantization on B200 or 2x H100 — a deployment profile that previously required either a closed API or a license-restricted open model. τ²-Bench Telecom jumped 37% to 85% generation-over-generation; AA-Omniscience Non-Hallucination #1 at 86%; 48 languages. Procurement teams running RAG or agentic workflows in regulated industries should evaluate this within the next two sprints; strongest non-Chinese open-weight frontier-adjacent candidate in the Western catalog.

SOURCE: COHERE.COM/BLOG/COMMAND-A-PLUS, VENTUREBEAT.COM

2026-05-22

RELEASE 02

Microsoft Research

SPECIALIST

MULTIMODAL

Fara1.5-27B

Teams building computer-use agents should switch to open-weight Fara1.5 before scaling — Microsoft's tiny model beats OpenAI Operator and Google Computer Use on live web tasks.

Fara1.5-27B scores 72% on Online-Mind2Web vs 58.3% for OpenAI Operator and 57.3% for Gemini 2.5 Computer Use; even the 9B variant beats them at 63.4%. Built on Qwen3.5, open weights on Azure AI Foundry. Collapses the cost structure for browser-agent deployments — operators currently paying per-task to closed providers can run an entire fleet on a single host. Microsoft's strategy reveal: in browser-agent specialty, scale is no longer the moat.

SOURCE: MICROSOFT.COM/EN-US/RESEARCH/ARTICLES/FARA1-5-COMPUTER-USE-AGENT, DECRYPT.CO

2026-05-21

RELEASE 03

Tencent

SPECIALIST MOE

Hunyuan Hy-MT2-30B-A3B

Localization and globalization ops teams should test Hy-MT2 against incumbent MT APIs — 30B MoE running 3B active now leads open-source translation; 1.8B variant runs offline on a phone.

Family release (1.8B / 7B / 30B-A3B) supporting 33 languages + 5 Chinese dialects. The 1.8B variant fits in 440MB after 1.25-bit AngelSlim quantization — runs offline on a phone. The 7B and 30B-A3B beat DeepSeek-V4-Pro and Kimi K2.6 on fast-thinking translation. Architects shipping multilingual products should benchmark: this is the first open-weight family where on-device offline translation is genuinely production-grade, not a research curiosity.

SOURCE: GITHUB.COM/TENCENT-HUNYUAN/HY-MT2

ARCHITECTURE WATCH

Patterns to track.

5 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

Speed-intelligence Pareto frontier reset by Flash tier

`Gemini 3.5 Flash (~280 tok/sec, AA Index 55+)`

`Command A+ (W4A4 quantized, 2x H100)`

`Gemini 3.1 Flash-Lite (carryover anchor)`

Gemini 3.5 Flash at 278-284 output tokens per second with AA Intelligence Index 55+ has redrawn the Pareto frontier — Flash-class models now match prior-generation Pro intelligence at sub-100ms-per-token throughput. Architects budgeting for agentic loops should recompute cost-per-task assuming Flash-tier models now do what required a Pro model six months ago. Closed Pro-tier models retain a 3-5 point intelligence lead but at 4-6x the latency.

SOURCE: ARTIFICIALANALYSIS.AI/ARTICLES/COHERE-LAUNCHES-OPEN-WEIGHTS-MODEL-COMMAND-A, DEEPMIND.GOOGLE/MODELS/GEMINI/FLASH

PATTERN 02

Agent-first ground-up redesigns replace chat-first architectures

`Gemini 3.5 Flash + Antigravity 2.0` `Qwen3.7-Max (35h autonomous demo)`

`Command A+ (unified four prior single-purpose models)`

`Mistral Medium 3.5 (anchor)`

Every flagship model launched this week was explicitly engineered for sustained multi-step agentic workloads — long-horizon planning, parallel tool use, persistent state — not for conversational turns. Antigravity 2.0 ships as a standalone agent-first GUI replacing the IDE paradigm; Command A+ collapsed four prior single-purpose models into one unified weights. Architects assuming a chat-first interaction model in their AI stack should review their roadmaps — vendor pricing, latency tuning, and API shapes are all moving toward agents as the unit of work.

SOURCE: ANTIGRAVITY.GOOGLE/BLOG/INTRODUCING-GOOGLE-ANTIGRAVITY-2-0, COHERE.COM/BLOG/COMMAND-A-PLUS

PATTERN 03

Open-weight frontier shifts from license-restricted to fully Apache 2.0

`Command A+ (Cohere Apache 2.0)` `Gemma 4 (Apache 2.0)` `DeepSeek V4 (MIT)` `Mistral Medium 3.5 (modified MIT)`

Cohere's Apache 2.0 release of a 218B MoE — its first ever under that license — joins Gemma 4's Apache 2.0 shift in marking a definitive move away from license-restricted open weights at the frontier. Llama Community License-style restrictions are now the laggard, not the standard. Procurement teams that previously had to lawyer up around 700M-MAU caps and field-of-use carveouts have a fast-growing set of frontier-adjacent options without that overhead.

SOURCE: VENTUREBEAT.COM/TECHNOLOGY/COHERE-CRACKS-LOSSLESS-QUANTIZATION-AND-NATIVE-CITATIONS-WITH-FIRST-FULL-APACHE-2-0-LICENSED-OPEN-MODEL-COMMAND-A

PATTERN 04

1M-token context becomes table stakes for closed frontier

`Gemini 3.5 Flash (1M)` `Qwen3.7-Max (1M)` `DeepSeek V4 Pro/Flash (1M)` `Claude Opus 4.7 (1M)` `GPT-5.5 Pro (1M)`

Every major frontier or frontier-adjacent model shipped this month carries a 1M-token context as the default, not a premium tier. The bar moved from 256K to 1M in two quarters. Architects building RAG or long-document workflows should stop treating retrieval chunking as the default — single-shot whole-corpus inference is now economically viable for many workloads. Mid-tier models still at 128K-256K (Command A+ at 128K input) are now the exception that needs justification.

SOURCE: BENCHLM.AI/BENCHMARKS/ARTIFICIALANALYSIS, MODEL CARDS ACROSS LABS

PATTERN 05

Specialist computer-use models challenge generalist agents on UI tasks

`Fara1.5-27B (72% Online-Mind2Web)` `Yutori Navigator n1 (64.7%)` `OpenAI Operator (58.3%)` `Gemini 2.5 Computer Use (57.3%)`

Fara1.5-27B's 72% Online-Mind2Web beats OpenAI's Operator (58.3%) and Gemini 2.5 Computer Use (57.3%) — purpose-built browser agents on fine-tuned small bases now outperform frontier generalists on web-task completion. Operators paying per-task on closed Operator / Computer Use APIs should run a TCO comparison against a self-hosted Fara1.5 fleet. The pattern: specialist fine-tunes on focused data beat broad-but-shallow frontier generalists on narrow real-world workflows.

SOURCE: DECRYPT.CO/368807/MICROSOFT-FARA15-OPEN-SOURCE-AI-BEATS-OPENAI-GEMINI, MICROSOFT.COM/EN-US/RESEARCH/ARTICLES/FARA1-5-COMPUTER-USE-AGENT

BENCHMARK MOVES

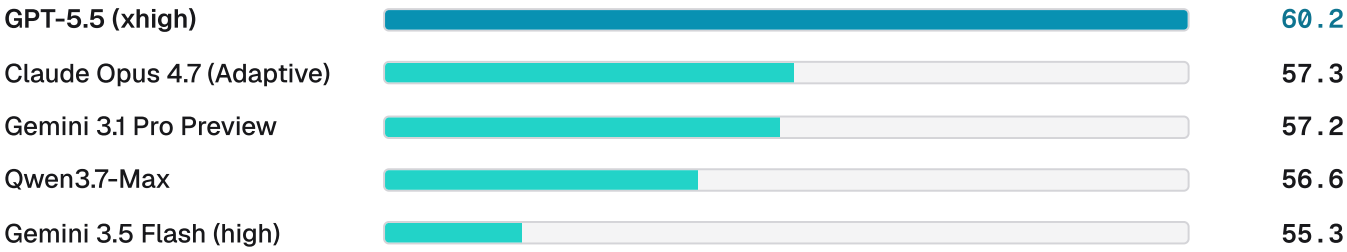
Where the leaderboard moved.

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

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX (SNAPSHOT MAY 21, 2026)

MOVE 01

Top remains GPT-5.5 (60.2); Qwen3.7-Max debuts at 56.6 — first Chinese model in global top 5, ahead of Gemini 3.5 Flash (55.3); Claude Opus 4.7 holds 57.3.

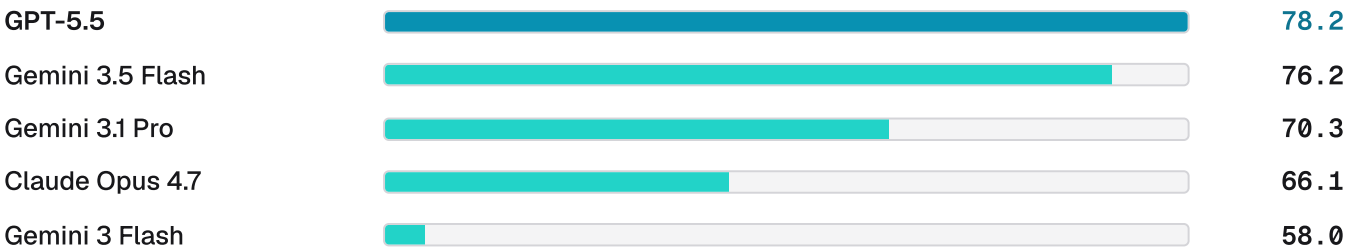


SOURCE: ARTIFICIALANALYSIS.AI/ARTICLES/COHERE-LAUNCHES-OPEN-WEIGHTS-MODEL-COMMAND-A

TERMINAL-BENCH 2.1 (AGENTIC TERMINAL CODING)

MOVE 02

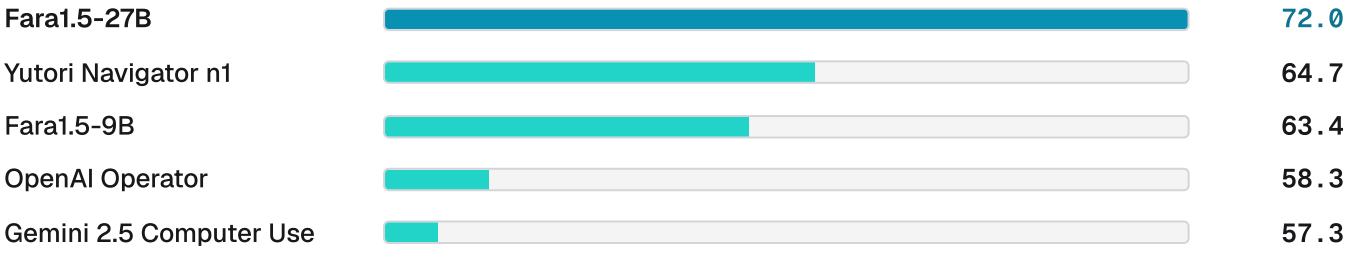
Gemini 3.5 Flash jumps to 76.2% — outscores Claude Opus 4.7 (66.1%), Gemini 3.1 Pro (70.3%), and Gemini 3 Flash (58.0%); GPT-5.5 leads at 78.2%.



SOURCE: DEEPMIND.GOOGLE/MODELS/GEMINI/FLASH

ONLINE-MIND2WEB (BROWSER COMPUTER-USE, 300 TASKS ACROSS 136 LIVE SITES) MOVE 03

Open-weight Fara1.5-27B sets new SOTA at 72.0%, beating OpenAI Operator (58.3%), Gemini 2.5 Computer Use (57.3%), and Yutori Navigator n1 (64.7%).



SOURCE: MICROSOFT.COM/EN-US/RESEARCH/ARTICLES/FARA1-5-COMPUTER-USE-AGENT

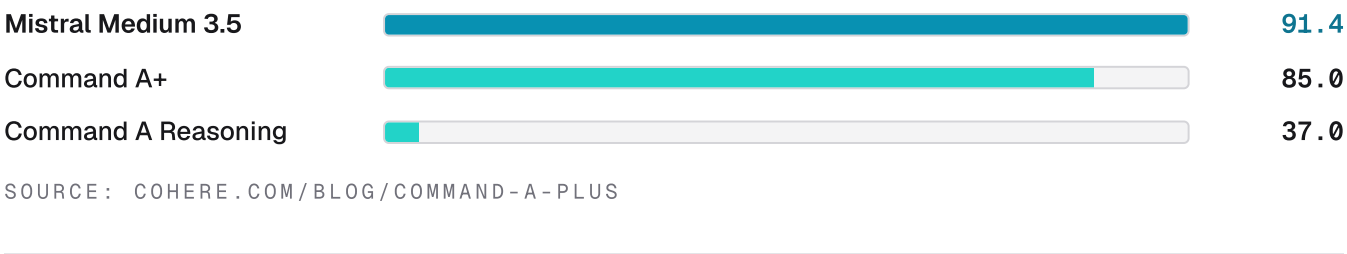
GDPVAL-AA (ECONOMICALLY VALUABLE KNOWLEDGE WORK, ELO) MOVE 04

Gemini 3.5 Flash reaches 1656 Elo — leapfrogging Gemini 3.1 Pro (1314) and Gemini 3 Flash (1204); Claude Opus 4.7 still leads enterprise reasoning at 1753.

SOURCE: DEEPMIND.GOOGLE/MODELS/GEMINI/FLASH

T²-BENCH TELECOM (AGENTIC TOOL USE) MOVE 05

Command A+ jumps from 37% (Command A Reasoning) to 85% — a 48-point generation-over-generation gain on open weights.



SOURCE: COHERE.COM/BLOG/COMMAND-A-PLUS

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of May 23, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**GPT-5.5 (xhigh) — AA Index 60.2**	Claude Opus 4.7 (Adaptive Max) — AA Index 57.3, LMArena 1492-1501	Headline intelligence crown unchanged; Gemini 3.5 Pro arrives next month and will be the real challenge to GPT-5.5's lead.
Open frontier	**Kimi K2.6 (Moonshot) — AA Index lead among open; DeepSeek V4 Pro #2 on open**	Command A+ (Cohere) — first Apache 2.0 frontier-adjacent MoE; Mistral Medium 3.5 dense flagship	Cohere made sovereign-AI on-prem deployment a real Western option this week; Chinese labs still hold raw intelligence lead.
Reasoning	**GPT-5.5 Pro / Claude Opus 4.7 (adaptive thinking)**	Qwen3.7-Max — 35-hour autonomous tool-use, AA 56.6 with strong long-horizon traces	Reasoning tier is now where China is closest to parity — buyers should test long-horizon agent loops side-by-side.
Coding	**Claude Opus 4.7 — SWE-Bench Verified 87.6%, SWE-Bench Pro 64.3%**	Gemini 3.5 Flash — agentic-coding Pareto leader at 4x speed (Terminal-Bench 76.2)	Opus 4.7 owns hard coding; Flash owns sustained agentic coding loops where token-per-second matters more than peak score.
Multimodal	**Gemini Omni Flash — native text/image/audio/video to grounded video**	Gemini 3.5 Flash on input multimodal (CharXiv 84.2, MMMU-Pro 83.6)	Google now owns both ends of multimodal — input understanding and conditional video output — without a clear cross-vendor competitor this week.
Edge / small	**Gemma 4 family (E2B / E4B / 26B-MoE / 31B) — Apache 2.0, runs from Pi to workstation**	Hy-MT2-1.8B (440MB at 1.25-bit) for translation; Fara1.5-4B for browser agents	Edge tier consolidates around specialists — translation, browser-use, multimodal — rather than general-purpose miniatures.

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.

2026-05-19

SIGNAL 01

Google

Gemini 3.5 Flash + Omni Flash + Antigravity 2.0 + Managed Agents API at I/O — three-month product cycle compressed into one keynote

Google shifted the conversation from 'model choice' to 'agent platform choice'. Procurement teams should reopen Gemini-vs-incumbent bake-offs that closed in Q1; the agent-harness lock-in (AGENTS.md, SKILL.md, managed Linux sandboxes) is now the strategic battleground, not the model itself.

SOURCE: BLOG.GOOGLE/INNOVATION-AND-AI/TECHNOLOGY/DEVELOPERS-TOOLS/GOOGLE-IO-2026-DEVELOPER-HIGHLIGHTS

2026-05-20

SIGNAL 02

NVIDIA

Q1 FY27 record $81.6B revenue (+85% YoY); Vera Rubin + Vera CPU + BlueField-4 STX announced; Vera CPUs already shipping to OpenAI, Oracle, Anthropic, xAI

Data Center revenue $75.2B (+92% YoY) confirms the buildout has not yet peaked; Vera Rubin Q3 2026 ship date with a purpose-built agentic-AI CPU reframes 2H training capacity. CFOs and infrastructure architects should assume frontier training compute remains supply-constrained through year-end and budget GPU access on Q3-availability cadence, not on price.

SOURCE: INVESTOR.NVIDIA.COM/NEWS/PRESS-RELEASE-DETAILS/2026/NVIDIA-ANNOUNCES-FINANCIAL-RESULTS-FOR-FIRST-QUARTER-FISCAL-2027

2026-05-19

SIGNAL 03

Anthropic

Claude Managed Agents add self-hosted sandboxes (public beta) and MCP tunnels (research preview) — agent execution moves into customer infrastructure boundary

Anthropic's response to I/O wasn't a new model — it was an enterprise security control surface. Security and platform teams should evaluate this as the new baseline for agent-deployment threat models: tool execution stays on-prem, credentials never enter agent context, only outbound traffic. Buyers that previously gated Claude agents on data-leakage risk now have a clear path forward.

SOURCE: CLAUDE.COM/FR/BLOG/CLAUDE-MANAGED-AGENTS-UPDATES

2026-05-21

SIGNAL 04

OpenAI

Codex 'Goal mode' GA, macOS Appshots, locked-Mac computer use, browser annotations + Dell hybrid deployment partnership (May 18-19)

OpenAI did not counter-launch a model into I/O week — instead doubled down on Codex as the front door for enterprise on-prem (Dell AI Factory) and pushed agent persistence (Goal mode, locked computer use). Engineering leaders evaluating coding-agent vendors should now treat Codex as a multi-surface platform (CLI + IDE + mobile + macOS app + Windows incoming) rather than a chat feature, and re-evaluate hybrid deployment paths.

SOURCE: HELP.OPENAI.COM/EN/ARTICLES/6825453-CHATGPT-RELEASE-NOTES

2026-05-20

SIGNAL 05

Alibaba

Alibaba Cloud Summit: Qwen3.7-Max + Qwen3.7-Plus-Preview + homegrown Zhenwu M890 AI chip (3x predecessor) + rebuilt full-stack agentic cloud platform

Alibaba projects model & application services ARR of ¥10B (~$1.4B) in the June quarter and ¥30B (~$4.1B) by year-end — a credible commercial AI business at scale. Buyers in APAC or with China-region data residency requirements should treat Alibaba as a primary frontier vendor evaluation this quarter, not a regional alternative.

SOURCE: ALIBABACLOUD.COM/BLOG/ALIBABA-UNVEILS-NEW-AI-CHIP-FLAGSHIP-MODEL-AND-REBUILT-CLOUD-STACK-AI-FOR-AGENTIC-ERA

2026-05-20

SIGNAL 06

Cohere

Command A+ shipped under Apache 2.0 — first ever fully-permissive license from Cohere on a frontier-adjacent MoE

Sovereign-AI and data-residency buyers now have a Western on-prem frontier-adjacent option without license restrictions. Strategic signal: Cohere is positioning as the Western counter to Chinese open-weights labs (DeepSeek, Moonshot, Z.ai) for regulated enterprises that cannot use Chinese weights. Procurement teams should test Command A+ vs Mistral Medium 3.5 for RAG and tool-use workloads inside the next two sprints.

SOURCE: COHERE.COM/BLOG/COMMAND-A-PLUS

WATCHLIST

On the radar next.

7 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 25 - 30

WATCH 01

Anthropic round close — final terms, lead investor, full investor list

Bloomberg May 22 reported close 'as soon as next week' at $30B-plus / $900B-plus. Resets the frontier-lab valuation curve 15x in 14 months. A slip past end-May or a haircut below $700B would be the loudest re-rating signal in a year.

MAY 27 - 28

WATCH 02

Samsung-union ratification vote on tentative HBM4 agreement

Binary outcome. Approval keeps the 18-day walkout off the table and HBM4 supply for Vera Rubin intact; rejection re-activates the W20 supply-risk thesis on three days' notice.

MAY 26 - JUNE 13

WATCH 03

Anthropic Opus 4.8 — leaked partner evaluations, expected I/O response

Multiple sources reporting select Anthropic partners running Opus 4.8 internal evals; Anthropic's I/O-week response was infrastructure (sandboxes, MCP tunnels), not a model. The model itself likely lands within 3 weeks based on Opus 4.6 to 4.7 cadence.

JUNE 2 - 3

WATCH 04

Microsoft Build 2026 (San Francisco) — MAI-line announcements, Foundry updates, Fara1.5 enterprise rollout

Build was pushed from May to June 2-3 — Microsoft has been notably quiet on its own MAI model line through I/O week. Build is where MAI frontier models, Copilot agent stack moves, and Foundry consolidation are likely to land; procurement teams running Azure / Foundry should hold model-selection decisions until after Build.

MAY 26 - JULY 15

WATCH 05

Gemini 3.5 Pro launch — frontier intelligence rematch

Google promised Pro 'next month' (June 2026). This is the model that determines whether Google retakes the AA Intelligence Index lead from GPT-5.5 (currently 60.2) or stays on the second podium tier. Buyers planning Pro-tier procurement should hold pending Pro pricing and benchmarks.

MAY 26 - JULY 31

WATCH 06

DeepSeek R3 / V4-Thinking — overdue per W20 watchlist

DeepSeek V4 Pro/Flash landed April 24 with strong benchmarks but explicitly without R3-style deep reasoning weights. Cohere CEO and DeepSeek itself have publicly acknowledged being 3-6 months behind US frontier — an R3 release is the catalyst that could close that gap or confirm it.

JUNE 9 - 13

WATCH 07

Apple WWDC 2026 — on-device model strategy, Apple Intelligence v2

Apple absent from the model-launch cycle through Q2; WWDC is the only major Apple model-disclosure window in 1H. Mobile / edge architects building on-device AI should expect significant Apple Intelligence repositioning and possibly partnered closed-model integration changes.

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 05

Week 21 of 2026 · May 23, 2026

WEB

brianletort.ai/industry/model

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