THE MODEL PULSE ISSUE 03 · WEEK 19 OF 2026

THE BIG READ

Frontier text went quiet. Voice crossed into chain-of-thought. The week's biggest capability move was a compute deal.

KEY SIGNALS THIS PERIOD

7	4	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 8, 2026

Issue 03 · 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.

W19 was the week the model layer split. Frontier text shifted into a maintenance cadence — no new flagship landed, the AA Intelligence Index ceiling held at GPT-5.5 (xhigh) = 60, and OpenAI's notable release was a tuned variant: GPT-5.5 Instant became the new ChatGPT default with 52.5% fewer hallucinations on high-stakes prompts. Google's Gemini 3.2 Flash surfaced in the iOS app and AI Studio without an announcement two weeks before I/O — leak-grade only. The center of gravity moved sideways into voice and into compute supply.

OpenAI's Realtime trio (GPT-Realtime-2 + Translate + Whisper) shipped on May 7 with GPT-5-class reasoning baked into the audio stack — Big Bench Audio jumped 15.2 points, Audio MultiChallenge instruction-following 13.8 points. Inworld shipped Realtime TTS-2 on the same day with conversational turn-awareness and emotional steering. 'Voice model' is now a frontier tier with its own benchmarks (Big Bench Audio, Audio MultiChallenge, Speech Arena Elo), not a checkbox feature.

The most consequential capability move was a compute deal. Anthropic + SpaceX (May 6) — 300+ MW at Colossus 1 in Memphis, ~220K NVIDIA GPUs within the month — paired with same-morning announcements that doubled Claude Code 5-hour rate limits, removed peak-hours throttling, and raised Opus API limits. The W19 read: the binding constraint on frontier deployment has flipped from model capability to power-and-GPU supply.

The open frontier hardened on efficiency, not absolute capability. Zyphra ZAYA1-8B (May 6) hits AIME 2025 91.9% with 700M active parameters and Markovian RSA test-time compute — and is the first frontier-class MoE trained end-to-end on AMD MI300X + Pensando + ROCm via IBM Cloud. Allen AI EMO (May 8) shows 12.5% of experts can carry near-full performance via emergent semantic specialization. Three independent labs in two weeks have shipped reasoning-credible MoE models with ≤3B active parameters — architects pricing 70B-class inference should re-validate against the sub-billion-active band before committing 2027 capacity.

For the May 12-22 window, the load-bearing catalysts are NVIDIA Q1 FY27 (May 20), Google I/O 2026 (May 19-20), Microsoft Build (May 19-22), the CoreWeave 10-Q (May 11-21), and the Anthropic round close.

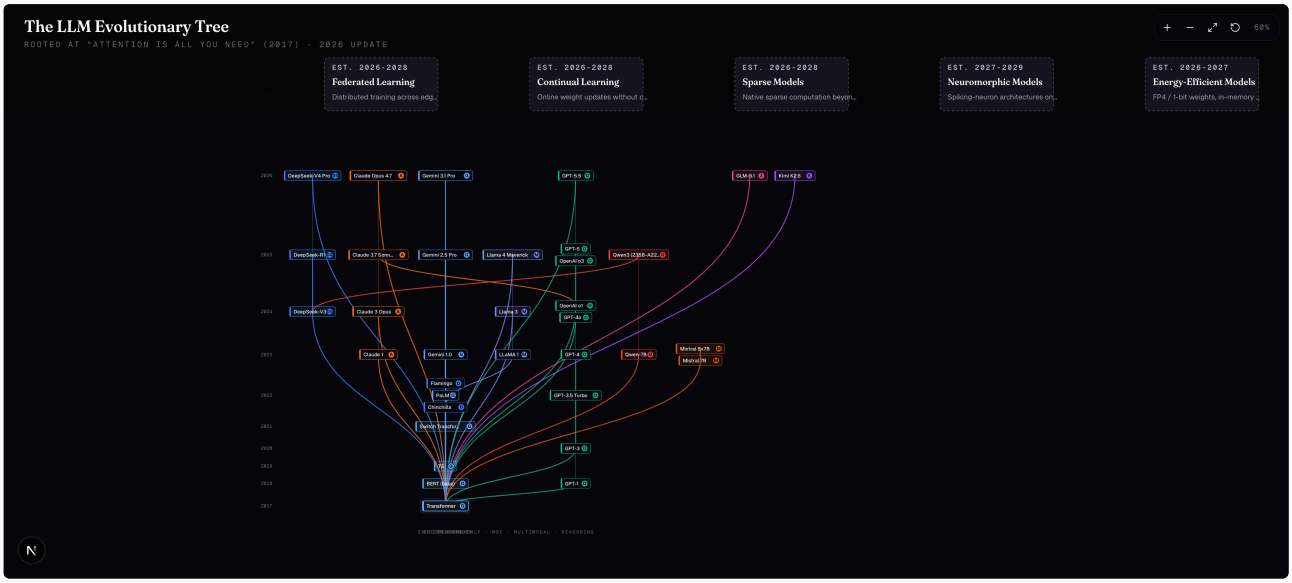
THE BAR

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.

7 model rows added across 4 vendors in a single 7-day window; the canopy widened in voice (OpenAI Realtime trio, Inworld TTS-2), efficient open-frontier MoE (Zyphra ZAYA1-8B), research-grade emergent expert specialization (Allen AI EMO), and a tuned ChatGPT default (GPT-5.5 Instant).



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

ADDED (7)

- gpt-5-5-instant
- gpt-realtime-2
- gpt-realtime-translate
- gpt-realtime-whisper
- inworld-realtime-tts-2
- zaya1-8b
- emo-1b14b

UPDATED (3)

- gpt-5-5
- claude-opus-4-7
- claude-mythos

Frontier text went into maintenance cadence: no new closed flagship, GPT-5.5 (xhigh) holds AA Index 60. The week's tree growth is in voice (OpenAI Realtime trio + Inworld TTS-2 ship on the same day with frontier-class scores) and at the efficient end of the open frontier (ZAYA1-8B at 700M active matches DeepSeek-R1; EMO at 1B active maintains near-full performance with 12.5% expert activity). Gemini 3.2 Flash leaked in iOS app + AI Studio metadata on May 5 but is held until Google I/O confirmation (May 19-20). Updated rows reflect non-release moves: GPT-5.5 acquires the Instant variant as default, Claude Opus 4.7 inherits the Anthropic ↔ SpaceX compute deal as a rate-limit boost rather than a model-weights change, Claude Mythos federal posture continues to thaw with White House drafted guidance and reported NSA usage.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-05-05

RELEASE 01

OpenAI

FRONTIER DENSE

GPT-5.5 Instant

GPT-5.5 Instant becomes the ChatGPT default with 52.5% fewer hallucinations on high-stakes prompts.

OpenAI is treating frontier text as a maintenance cadence: rather than ship a new flagship, they swapped the free-tier default to a tuned variant that materially reduces high-stakes hallucinations (medicine, law, finance) and tightens responses ~30% in word count. For boards and operators, the operational profile of free-tier ChatGPT just changed under-foot — the 'cheap fallback' tier now hits AIME 2025 81.2 and MMMU-Pro 76.0, not 65.4 — without any procurement event. Architects who routed low-stakes traffic to a non-frontier tier should re-evaluate the cost-per-quality bake-off.

SOURCE: OPENAI PRIMARY; THE VERGE; MASHABLE; ANDROID AUTHORITY

2026-05-07

RELEASE 02

OpenAI

FRONTIER MULTIMODAL

GPT-Realtime-2

First voice model with GPT-5-class reasoning, 128K context, adjustable effort — Big Bench Audio 96.6%.

OpenAI's Realtime API exits beta with a trio (Realtime-2 + Translate + Whisper). Realtime-2 jumps 15.2 points on Big Bench Audio (96.6% vs 81.4% for v1.5) and 13.8 points on Audio MultiChallenge instruction-following; Zillow reported call-success climbing from 69% to 95% on its hardest adversarial set. For boards and architects, voice/IVR replacement is now a 90-day procurement decision — the gap between a frontier voice model and an enterprise IVR is large enough to justify same-quarter pilots, especially with named enterprise references (Zillow, Priceline, Deutsche Telekom) and pricing at $32 / $64 per Mtok audio.

SOURCE: OPENAI PRIMARY; MARKTECHPOST; THE NEXT WEB; LATENT.SPACE

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.

2026-05-06

RELEASE 01

Zyphra

OPEN FRONTIER

MOE

ZAYA1-8B

8B-total / 700M-active MoE reasoning model, AMD-trained end-to-end, matches DeepSeek-R1 at sub-1B active.

Two procurement signals in one release. (1) The 'efficient open frontier' line from W18 (Laguna XS.2, OpenAI Privacy Filter) extends down to 700M active parameters with a Markovian Recursive Self-Attention test-time compute trick that pulls AIME 2025 to 91.9 — architects should treat 700M active as the new floor for serious reasoning workloads. (2) ZAYA1 is the first large-scale MoE foundation model trained on a full-stack AMD platform (MI300X + Pensando + ROCm via IBM Cloud) — an externally-verifiable proof point that AMD is now a credible frontier-training substrate, not just an inference option. Operators with 'sovereign / regulated / non-NVIDIA' constraints have a real second supply path with a benchmark trail.

SOURCE: ZYPHRA PRIMARY; ARXIV 2605.05365; VENTUREBEAT; HUGGING FACE

2026-05-05

RELEASE 02

Inworld

SPECIALIST

MULTIMODAL

Inworld Realtime TTS-2

Frontier-positioned voice model with conversational turn-awareness, emotional steering, one voice across 100+ languages.

Closed-weights but research-preview-priced via API, so listed as specialist alongside the OpenAI Realtime trio. Inworld already holds #1 on the Artificial Analysis Speech Arena (Elo 1209.6, 73.3% win rate); TTS-2 shifts the same line from voiceover/audiobook posture to real-time dialog with prior-turn context and natural-language steering ('[whisper]', '[hold back rage]'). Pairs with the OpenAI Realtime trio to make voice 'frontier' a procurement category this week. CISOs should evaluate prompt-injection and identity-spoofing controls before deploying directive-driven voice agents in customer-facing channels.

SOURCE: INWORLD PRIMARY; BUSINESSWIRE; LAS VEGAS SUN; INWORLD DOCS

2026-05-08

RELEASE 03

Allen AI (Ai2)

SPECIALIST MOE

EMO (Emergent Mixture-of-Experts)

14B / 1B-active research MoE where modular structure emerges from data — 12.5% of experts can carry near-full performance.

Research-grade rather than production, but architecturally important: by constraining same-document tokens to share an expert pool, EMO produces experts that specialize at the semantic level (math, code, biomedical) rather than the syntactic level (punctuation). Practical implication for serving cost: ~8x reduction by routing only the 8 of 128 experts a given workload needs, with 3pp accuracy cost. This is the line of research that will set the next 12 months of MoE serving economics — operators sizing 2027 inference clusters should track the emergent-specialization literature alongside parameter-count benchmarks.

SOURCE: ALLEN AI PRIMARY; ARXIV 2605.06663; HUGGING FACE

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

Voice models cross into chain-of-thought territory

`GPT-Realtime-2 (GPT-5-class reasoning baked in, 96.6% Big Bench Audio)`

`Inworld Realtime TTS-2 (turn-aware emotional reasoning)`

`GPT-Realtime-Translate (70+ input → 13 output languages)`

Two independent vendors shipped 'frontier voice' the same week, and both ship reasoning into the audio stack rather than treating speech as a thin wrapper around a text model. GPT-Realtime-2 exposes the same minimal/low/medium/high/xhigh effort knobs as GPT-5.5; Inworld TTS-2 adds turn-aware emotional reasoning. Procurement implication: 'voice model' is now a tier with its own benchmarks (Big Bench Audio, Audio MultiChallenge, Speech Arena Elo), not a checkbox feature — voice agent SLAs should be evaluated against 250+ turn benchmarks now, not single-prompt audio quality.

SOURCE: OPENAI PRIMARY; INWORLD PRIMARY; MARKTECHPOST; LATENT.SPACE

PATTERN 02

Sub-billion-active-param frontier, doubled

`ZAYA1-8B (700M active)` `EMO 1B14B (1B active, 8 of 128 experts)`

`Laguna XS.2 (3B active, W18)`

Three independent labs in two weeks have shipped reasoning-credible MoE models with ≤3B active parameters. ZAYA1 with test-time compute reaches AIME 2025 91.9; EMO can serve at 12.5% experts active with only 3pp drop. The cost curve for 'good enough on hard reasoning' has bent another time — architects pricing 70B-class inference should re-validate against this band before committing 2027 capacity, and CFOs should treat 2026 inference TCO models as already stale.

SOURCE: ZYPHRA PRIMARY; ALLEN AI PRIMARY; HUGGING FACE; VENTUREBEAT

PATTERN 03

Compressed and hybrid attention as the KV-cache default

`ZAYA1-8B (Compressed Convolutional Group-Query Attention, 8x KV reduction)`

`DeepSeek V4 Pro (hybrid local + long-range, transformers v5.8.0 support landed May 5)`

`MiMo V2.5 Pro (W18 — SWA + Global, 6:1)`

The Q1-2026 'pure attention' design is being squeezed out for hybrid forms that compress or shard the KV cache. For long-context serving, this changes the unit economics: ZAYA1's CCGQA reports 1.7x prefill latency improvement at 16K context on H100. Operators should expect new model lines this quarter to stop quoting full multi-head numbers as the default; KV-cache budget is now a first-class procurement variable.

SOURCE: ZYPHRA PRIMARY; ARXIV 2510.04476; HUGGINGFACE TRANSFORMERS V5.8.0 RELEASE NOTES

PATTERN 04

Compute-supply deals are the new capability lever

`Anthropic ↔ SpaceX (300 MW Colossus 1, ~220K NVIDIA GPUs within the month)`

`Rackspace ↔ AMD (governed AI Cloud MOU, May 7)`

`OpenAI on AWS Bedrock (Forbes coverage extends procurement narrative May 6)`

When the most consequential 'capability' move of the week — Claude Code rate limits doubled, peak-hours throttle removed — comes from a megawatt deal rather than a model release, the constraint stack has flipped. For boards and operators, the next 6 months of frontier deployment will be paced by power and silicon allocation, not parameter counts. Watch for federal customers to start asking vendors which compute partners back which SKUs, not which model version they're running.

SOURCE: ANTHROPIC PRIMARY; BLOOMBERG; CNBC; BUSINESS INSIDER; STOCKTITAN; FORBES

PATTERN 05

AMD-trained frontier crosses verification threshold

`ZAYA1-8B (full AMD MI300X + Pensando + ROCm stack via IBM Cloud)`

`Rackspace ↔ AMD MOU for governed Enterprise AI Cloud (May 7)`

`AMD Q1 2026 8-K (May 5) — DC revenue $5.8B (+57% YoY), Helios + Meta 6 GW shipping H2`

ZAYA1 is the first large-scale MoE foundation model with a published technical report claiming end-to-end AMD training; Rackspace + AMD's MOU lands the same week with governed-cloud framing for regulated workloads; AMD's Q1 print confirms a credible second supply path with Meta's 6 GW Helios commitment. Three independent W19 signals. AMD-as-training-substrate is no longer an aspirational story — it's a procurement option with a benchmark trail.

SOURCE: ZYPHRA PRIMARY; IBM CLOUD; STOCKTITAN/RACKSPACE; AMD 8-K

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.

BIG BENCH AUDIO (W19 VOICE-REASONING SHIFT) MOVE 01

GPT-Realtime-2 (high effort) jumps 15.2pp over v1.5; voice models close the speech-reasoning gap to text.

SOURCE: OPENAI PRIMARY; LATENT.SPACE; MARKTECHPOST

AUDIO MULTICHALLENGE (INSTRUCTION FOLLOWING ON SPEECH) MOVE 02

GPT-Realtime-2 xhigh hits 48.5%, +13.8pp over v1.5; Scale-AI independent eval shows instruction retention 36.7% → 70.8%.

SOURCE: OPENAI PRIMARY; MARKTECHPOST; THE NEXT WEB

AIME 2025 / HMMT 2025 (OPEN-WEIGHTS REASONING AT <1B ACTIVE) MOVE 03

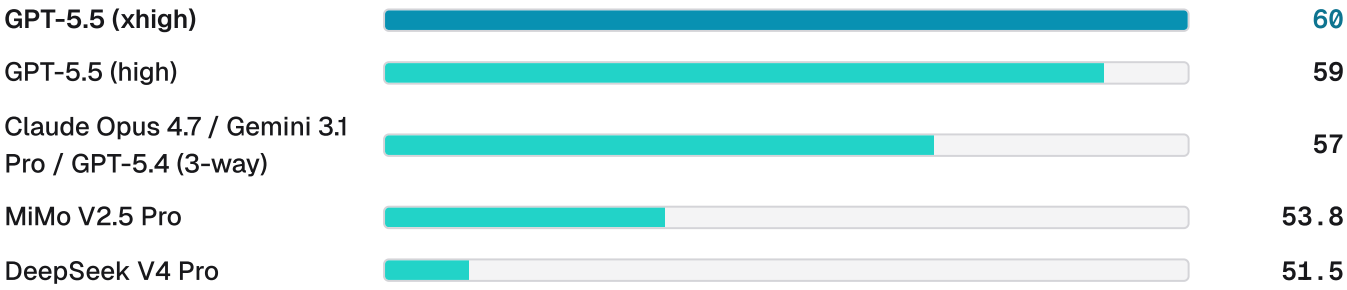
ZAYA1-8B (700M active) with test-time compute reaches frontier-grade math — narrows gap to GPT-5-High and Gemini-2.5 Pro.

ZAYA1-8B + Markovian RSA	AIME 91.9 / HMMT 89.6
ZAYA1-8B (single rollout)	Matches DeepSeek-R1-0528
Qwen3-4B / Gemma3-12B reference	Comparable

SOURCE: ZYPHRA PRIMARY; ARXIV 2605.05365; VENTUREBEAT

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX (W19 SNAPSHOT) MOVE 04

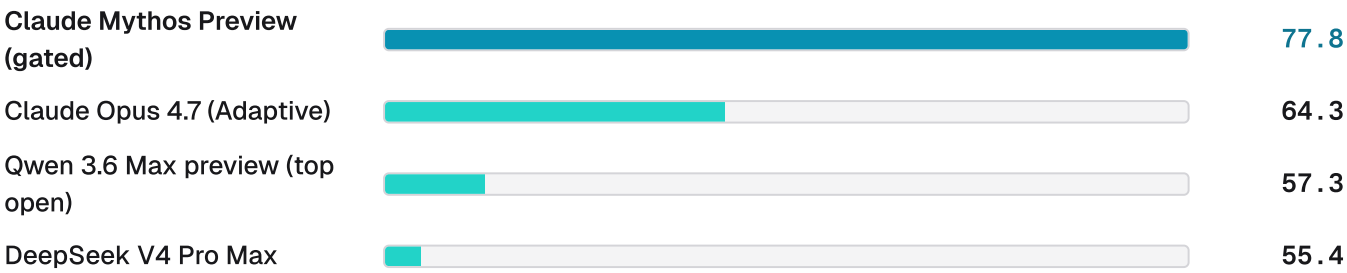
No leader change vs W18; the W18 picture (GPT-5.5 xhigh = 60) holds. Worth flagging — silence is the signal: frontier text is in maintenance.



SOURCE: ARTIFICIAL ANALYSIS PRIMARY; LMSPEED LEADERBOARD MIRROR (MAY 7 UPDATE)

SWE-BENCH PRO (W19 LEADERBOARD PULL, MAY 7) MOVE 05

Closed leaderboard ceiling unchanged; production-tier gap to top open weights narrowed to ~7pp.



SOURCE: BENCHLM LEADERBOARD PULL MAY 7; ARTIFICIAL ANALYSIS

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of May 8, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**OpenAI GPT-5.5 (xhigh) — AA Index 60**	Anthropic Claude Opus 4.7 / Google Gemini 3.1 Pro Preview (tied 57)	Unchanged from W18. The W19 sub-event is GPT-5.5 Instant becoming ChatGPT default (May 5) and an unconfirmed Gemini 3.2 Flash leak (May 5). No flagship change.
Open frontier	**Moonshot Kimi K2.6 / Xiaomi MiMo V2.5 Pro (tied AA ~54)**	DeepSeek V4 Pro (51.5)	Unchanged at the top. W19 movement is at the efficient end: ZAYA1-8B and EMO push 'sub-billion-active reasoning' as a credible procurement band.
Reasoning	**Anthropic Claude Mythos Preview (gated, 77.8 SWE-Bench Pro)**	OpenAI GPT-5.5 (xhigh, AA 60); Zyphra ZAYA1-8B with test-time compute (AIME 2025 91.9 at 700M active)	Mythos still gated behind Project Glasswing; W19 federal posture warming further (White House guidance, NSA usage). ZAYA1 enters as the 'reasoning-at-tiny-active-params' challenger.
Coding	**Anthropic Claude Opus 4.7 (64.3 SWE-Bench Pro)**	OpenAI GPT-5.5 (Codex CLI on Terminal-Bench 2.0 = 82.0, leads composite)	Unchanged. W19 procurement signal: Codex on Bedrock + Bedrock Managed Agents in limited preview started April 28; Forbes coverage on May 6 lands it in W19's federal/enterprise narrative.
Multimodal	**OpenAI GPT-Realtime-2 (voice frontier, May 7) + Google Gemini 3.1 Pro Preview (text/image)**	Inworld Realtime TTS-2 (voice, #1 Speech Arena Elo); NVIDIA Nemotron 3 Nano Omni (open multimodal, W18)	Refreshed. Voice/realtime is the W19 multimodal story: GPT-Realtime-2 + Inworld TTS-2 both ship 'frontier-class' voice the same week. Gemini 3.2 Flash leak hints at multimodal price-band compression but is unconfirmed.
Edge / small	**Zyphra ZAYA1-8B (8B / 700M active, AMD-trained, AIME 2025 91.9)**	Poolside Laguna XS.2 (33B / 3B active, W18); Allen AI EMO (1B active research); OpenAI Privacy Filter (1.5B / 50M, W18)	Refreshed. ZAYA1 enters at the very bottom of the active-param band with reasoning-credible scores; AMD-trained, which makes it a substrate signal too. EMO reinforces the band as a research-grade 1B-active design.

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.

2026-05-06

SIGNAL 01

Anthropic + SpaceX

300 MW / 220K-NVIDIA-GPU compute deal at Colossus 1 (Memphis); Claude Code 5-hour rate limits doubled, peak-hours throttle removed for Pro/Max, Opus API limits raised; Claude Managed Agents ships Multiagent Orchestration, Outcomes, and Dreaming.

The W19 'capability' event for Claude was a power-and-silicon deal, not a model. For procurement, this is a clean read: Anthropic's binding constraint was compute, not training, and the throttle relief is downstream of capacity arriving 'within one month.' Architects locked into Claude Code can plan against materially higher concurrent throughput before mid-summer; CISOs should request Anthropic's Multiagent Orchestration security model before deploying agent-coordination patterns.

SOURCE: ANTHROPIC PRIMARY; BLOOMBERG; CNBC; BUSINESS INSIDER; FRANCE24; CBC; ARS TECHNICA

2026-05-06

SIGNAL 02

OpenAI on AWS Bedrock

GPT-5.5 + Codex + Bedrock Managed Agents pick up procurement-press traction (Forbes); limited preview started April 28, but federal/enterprise procurement coverage compounded in W19.

Federal and AWS-anchored enterprise teams can now consume OpenAI frontier through existing Bedrock IAM, PrivateLink, KMS, and CloudTrail with usage applied against AWS commitments. The single biggest 'is OpenAI procurable in our existing posture?' objection just dropped. Closes the W18 watchlist item explicitly; watch AWS customer-disclosure announcements before the limited preview hardens at GA.

SOURCE: FORBES (MARKMAN) 2026-05-06; AWS WHAT'S NEW 2026-04-28; OPENAI PRIMARY; ABOUT AMAZON; SILICONANGLE

2026-05-05

SIGNAL 03

US CAISI / Google DeepMind / Microsoft / xAI

US Center for AI Standards and Innovation (Department of Commerce) signs pre-deployment evaluation agreements with Google DeepMind, Microsoft, and xAI — completing perimeter coverage of the top-five US labs after renegotiated agreements with OpenAI and Anthropic. Labs will provide early access including model versions 'with lowered safety protections' for national-security testing in controlled settings.

Every US frontier model now ships through a federal capability/risk gate before public release. Vendor-risk frameworks need a CAISI-status dimension alongside availability SLAs. For boards, this is the new cost of doing frontier business in the US — and a leading indicator that international frontier labs (Mistral, Alibaba, DeepSeek, Moonshot, Xiaomi) face an asymmetric procurement read against the same federal customers.

SOURCE: BBC NEWS; CNBC; THE HILL; NEURA.MARKET

2026-05-06

SIGNAL 04

Scale AI / DoD CDAO

Scale AI awarded $500M Pentagon contract (5x prior $100M deal) for AI integration into military decision-making and data-processing workflows.

Federal AI procurement continues to consolidate around 'data-and-eval' middleware vendors, not just model labs. For investors and operators tracking the model layer, this is the procurement-stack signal: model access (Bedrock-OpenAI, Anthropic-GSA-$1, DoD eight-vendor) is being matched by data-and-eval scaffolding contracts at increasing magnitude.

SOURCE: THE NEXT WEB 2026-05-06

2026-05-07

SIGNAL 05

Rackspace + AMD

MOU for governed Enterprise AI Cloud — AMD Instinct GPUs + EPYC CPUs in a managed stack with private/hybrid posture, Enterprise Inference Engine, dedicated bare-metal SLAs.

Combined with ZAYA1 the same week, AMD now has a credible training-substrate proof point AND a regulated-workload sales motion. Architects with sovereign / regulated / non-NVIDIA constraints have a real second supply path with named commercial packaging, not just a marketing one.

SOURCE: STOCKTITAN / RACKSPACE 2026-05-07

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

Likely venue for Gemini 3.2 Flash confirmation (the May 5 leak surfaced in iOS app + AI Studio metadata at $0.25 / $2.00 per Mtok) and any Gemini 3.2 Pro announcement that would force a frontier scorecard refresh. Watch for fabric-side disclosures: Google has been silent through W18-W19 while OpenAI / Microsoft / NVIDIA shipped MRC publicly.

MAY 12-22

WATCH 02

Anthropic round close

Talks have been 'within two weeks' of close for two weeks. The price ($850B floor / $900B mid / $1T per FT) and lead-investor disclosure re-rate the entire frontier-lab valuation curve. A walk-away from the price would be the loudest re-rating signal in a year; a clean close at $1T validates the FT signal as primary.

MAY 11-21

WATCH 03

CoreWeave 10-Q filing window

Audited detail behind the W19 Q1 print: $99.4B backlog conversion schedule, $31-35B 2026 capex composition, customer concentration, NVIDIA $2B Class A equity terms. The ~36% expected to recognize within 24 months is the line short-side analysts will read first; material differences with management commentary would re-rate the entire neocloud category.

MAY 19-22

WATCH 04

Microsoft Build 2026

Azure AI roadmap, Maia 200 numbers, Copilot Studio agent-runtime updates, and any production MRC fabric disclosure from Microsoft Fairwater are all on the agenda. A first-disclosure of an MRC fabric at >50,000-GPU scale outside Microsoft would test prediction p13.

MAY 5-22

WATCH 05

DeepSeek-R3 reasoning model release

DeepSeek shipped V4 Pro and V4 Flash on Apr 24 but the R-series reasoning sibling has not landed for V4. Given the 8-week R1 → R2 cadence, R3 is overdue; if it lands within the watchlist window it likely closes the AA Intelligence Index gap to GPT-5.5 below 4 points.

LATE MAY 2026

WATCH 06

xAI Grok 4.4 cadence

xAI moved Grok 4.3 to GA on May 6. xAI has been on a ~3-week cadence for Grok 4.x. A Grok 4.4 within the watchlist window would close the AA Intelligence Index gap to Opus 4.7 (currently 4 points); separately, watch for the first xAI multimodal / robotics model crossover with Tesla FSD given the SpaceX Terafab + Tesla joint-fab announcement (May 6).

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 03

Week 19 of 2026 · May 8, 2026

WEB

brianletort.ai/industry/model

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