AI News

Trends — April 2026

2026-05-02 - 14:22

Compiled with Claude AI

1 Executive Summary

April 2026 produced a dense sequence of *frontier model* releases. **Anthropic** shipped *Claude* Opus 4.7 on the 16th and confirmed but withheld *Claude* Mythos on the 7th, citing internal AI Safety Levels protocols (*ASL-4*) triggered for the first time in commercial AI history. **OpenAI** released *GPT*-5.5 on the 23rd, six weeks after *GPT*-5.4. **Google** released the Gemma 4 family under Apache 2.0 on the 2nd. **Meta** shipped Llama 4 Scout and Maverick on the 5th and reversed its open-weights posture three days later with the proprietary Muse Spark. The closing week added **Moonshot AI** Kimi K2.6, **DeepSeek** V4 at 1.6 trillion *parameters*, and Grok 4.3 from **xAI** within days of each other.

Coding capability was the focal axis of competition. Opus 4.7 set new highs on *SWE-bench* Verified (87.6%) and *SWE-bench* Pro (64.3%); *GPT*-5.5 led *Terminal-Bench 2.0* (82.7%) and *OSWorld-Verified* (78.7%). Reasoning benchmarks such as *GPQA Diamond* have effectively saturated at the frontier — Opus 4.7, *GPT*-5.5, and Gemini 3.1 Pro all sit within 0.4 points of each other.

Industry context was dominated by the closing of Q1 2026 venture data: $297 billion deployed globally, with AI-related companies absorbing 81%. **SpaceX** completed the acquisition of **xAI** on terms valuing the combined entity at $1.25 trillion. The European Commission moved closer to enforcement of the *EU AI Act*, with the *Digital Omnibus on AI* still under negotiation between Parliament and Council. Compute infrastructure constraints — measured in gigawatts of grid demand rather than dollars — became a structural concern that frontier labs and *hyperscalers* addressed openly for the first time.

2 Large Language Models

Eight frontier or near-frontier models reached general availability or limited preview in April. The cadence is a continuation of the pattern established in Q1, in which release windows shortened to roughly six weeks at **OpenAI** and roughly two months at **Anthropic**, with Chinese labs and **Meta** releasing on overlapping but independent schedules.

2.1 Releases and updates

Google released the Gemma 4 family on April 2 under the Apache 2.0 license. The family contains four variants ranging from 2.3 to 31 billion *parameters*, all natively *multimodal* across text, image, and video; the two larger variants additionally support audio. The 31B Dense variant ranks third globally on Arena AI among open models. Distribution from day one included **Hugging Face**, Ollama, Kaggle, and **Google** AI Studio.

Meta released Llama 4 Scout and Llama 4 Maverick on April 5. Both are *open-weight Mixture of Experts* models trained with native multimodality from pretraining rather than via later adapter integration. Scout offers a 10-million-*token context window*; Maverick a one-million-*token* window. Three days later, on April 8, **Meta** released Muse Spark — the company's first proprietary

model. Muse Spark scores 52 on the *Artificial Analysis Intelligence Index*. The proprietary release marked an explicit reversal of the open-weights posture that had defined **Meta**'s AI strategy through the Llama family.

Anthropic confirmed the existence of *Claude* Mythos on April 7 under what the company calls *Project Glasswing* — an access programme limited to roughly fifty enterprise partners. Mythos scored 93.9% on *SWE-bench* Verified and 94.6% on *GPQA Diamond*, both above any publicly released model. **Anthropic** chose not to release Mythos to general availability, citing the model's autonomous identification of thousands of *zero-day* vulnerabilities during testing as triggering *AI Safety Level* 4 protocols. The company stated the model would not be made generally available in the near term.

On April 16 **Anthropic** released *Claude* Opus 4.7 to general availability across *Claude* products, the *API* (*claude*-opus-4-7), **Amazon** Bedrock, **Google** Cloud Vertex AI, **Microsoft** Foundry, and **GitHub** Copilot Pro+, Business, and Enterprise tiers. The release continued **Anthropic**'s roughly two-month cadence following Opus 4.6 in February.

OpenAI released *GPT*-5.5 on April 23, internal codename Spud. The release was **OpenAI**'s first fully retrained base model since *GPT*-4.5 and is natively omnimodal — text, image, audio, and video processed through a unified architecture rather than stitched encoders. *API* access followed one day later on April 24. *GPT*-5.5 ships in three consumer surfaces: default *GPT*-5.5, *GPT*-5.5 Thinking, and *GPT*-5.5 Pro.

The closing week produced near-simultaneous releases from **Moonshot AI** (Kimi K2.6), **DeepSeek** (V4 at 1.6 trillion *parameters*, the largest publicly announced model), and **xAI** (Grok 4.3, promoted out of preview). **Alibaba** shipped Qwen 3.6-Plus, **Zhipu AI** released GLM-5.1 under MIT license, and **Arcee** released Trinity at 400 billion *parameters*.

2.2 Benchmarks and capabilities

Coding capability was the principal axis of differentiation in April. *Claude* Opus 4.7 scored 87.6% on *SWE-bench* Verified, up from 80.8% on Opus 4.6 and ahead of *GPT*-5.4 at 82.0% and Gemini 3.1 Pro at 80.6%. On the harder *SWE-bench* Pro, Opus 4.7 reached 64.3%, leading *GPT*-5.4 (57.7%) and Gemini 3.1 Pro (54.2%). On *CursorBench*, an evaluation produced by the **Cursor** IDE on real developer workflows, Opus 4.7 jumped from 58% to 70%.

GPT-5.5 led different coding-adjacent benchmarks. *Terminal-Bench 2.0*, which evaluates command-line task execution, recorded 82.7% for *GPT*-5.5 against 69.4% for Opus 4.7 and 68.5% for Gemini 3.1 Pro. *OSWorld-Verified*, which measures autonomous desktop task completion, showed *GPT*-5.5 at 78.7% and Opus 4.7 at 78.0%; both pass the 72.4% human expert baseline.

Reasoning benchmarks showed convergence. On *GPQA Diamond*, Opus 4.7 scored 94.2%, *GPT*-5.5 scored within the same range, and Gemini 3.1 Pro Preview reached 94.3%. Mythos in restricted access leads at 94.6%. The narrowness of these gaps — within a single percentage point at the top — has effectively saturated the *benchmark* for the publicly available frontier.

Long context capability progressed. Gemini 3.1 Ultra entered preview in April with a 2-million-*token context window* operating natively across text, image, audio, and video. *Claude* Opus 4.7 and *GPT*-5.5 each retain a one-million-*token* input context. Llama 4 Scout reached 10 million *tokens* — the largest published *context window* for an *open-weight* model.

Mythos's withholding from general release introduces a new pattern. *Claude* Mythos is the first model from a major lab to be completed and then withheld on safety grounds. The decision turned on the model's demonstrated ability to identify *zero-day* vulnerabilities autonomously during red-team testing, which **Anthropic** determined exceeded the controls applicable to general public access. Whether other labs adopt similar self-imposed restraint remains to be seen.

2.3 Pricing and access

Claude Opus 4.7 holds Opus 4.6's pricing of $5 per million input *tokens* and $25 per million output *tokens*. The new tokenizer encodes the same text into 1.0 to 1.35 times more *tokens* than Opus 4.6, so equivalent *prompts* incur a higher dollar cost despite identical per-*token* rates. The model is available across **Anthropic**'s full distribution: *claude*.ai (Pro, Max, Team, Enterprise), the *Claude API*, **Amazon** Bedrock, **Google** Cloud Vertex AI, **Microsoft** Foundry, and **Snowflake** Cortex AI. **GitHub** Copilot rolled out Opus 4.7 with a 7.5x premium multiplier as part of promotional pricing valid until April 30.

GPT-5.5's *API* pricing doubled relative to *GPT*-5.4: $5 per million input *tokens* and $30 per million output *tokens*, against $2.50 and $15 previously. *GPT*-5.5 Pro held at $30 and $180, unchanged from *GPT*-5.4 Pro. **OpenAI** states that *GPT*-5.5 uses approximately 40% fewer output *tokens* than *GPT*-5.4 on equivalent Codex tasks, partly offsetting the per-*token* increase. Batch and Flex pricing is half the standard rate; Priority pricing is 2.5x. On April 9 **OpenAI** also introduced a $100 ChatGPT Pro tier between Plus and the existing $200 Pro, the first subdivision of the Pro tier.

Gemini 3.1 Pro retains its undercut position at approximately $2 per million input and $12 per million output *tokens*, alongside its 2-million-*token* preview context. *Open-weight* options expanded materially in April — Gemma 4, Llama 4, GLM-5.1, Qwen 3.6-Plus, and **DeepSeek** V4 are downloadable for self-hosting at no licence fee. Arena AI and other independent leaderboards now place several open models within ten percentage points of the closed frontier on standard benchmarks.

3 Generative Media

April activity in generative media was concentrated in distribution and integration rather than in flagship-model releases. The pattern of recent quarters — model capability advancing fastest in editor-integrated workflows rather than in standalone *text-to-image*, *text-to-video*, or text-to-music interfaces — continued.

3.1 Image

OpenAI launched ChatGPT Images 2.0 on April 21, replacing *GPT* Image 1.5. The new system supports a Thinking Mode that integrates reasoning with web search and adds consistency across up to eight images per *prompt*. Distribution covers ChatGPT Plus, Pro, Business, and Enterprise; free-tier users remain on the previous version.

Black Forest Labs, **Stability AI**, and Midjourney did not release new flagship image models in April. The image generation segment has shifted toward integration with editor environments rather than standalone model launches; image work increasingly enters workflows through ChatGPT, **Google** Vids, and **Adobe** Firefly rather than through dedicated *text-to-image* interfaces.

3.2 Video

Google Vids gained native video generation through Veo 3.1 on April 2, available to anyone with a **Google** account at the rate of ten free generations per month. **Google** AI Pro and Ultra subscribers receive higher quotas, with up to one thousand Veo generations monthly available to Ultra accounts. The same release added a Chrome extension for screen recording and direct YouTube publishing.

Runway, **Pika Labs**, **Kling AI**, and **Luma Labs** continued iterative product updates in April without flagship-level releases. The market for AI video generation has consolidated around three high-end systems — **OpenAI** Sora 2.0, **Runway** Gen-4, and Veo 3.1 — with Pika and **Kling AI** competing on accessibility and music-aware workflows.

3.3 Music and audio

Google released Lyria 3 and Lyria 3 Pro inside **Google** Vids on April 2 as the music-generation counterpart to the Veo release. Custom music generation is gated to **Google** AI Pro and Ultra subscribers. **Suno** and **Udio** did not release new flagship models in April. **Udio** remained constrained by the October 2025 Universal Music Group settlement, which limits paid users to streaming-only access without download capability. **ElevenLabs** continued integration partnerships, including the LTX Studio audio-to-video pipeline launched in January.

4 Coding and Developer Tools

The April releases of *Claude* Opus 4.7 and *GPT*-5.5 were both positioned by their developers around coding and *agentic* workflows rather than chat or general

reasoning. The framing reflects where revenue concentration has gone: **Cursor** at $2 billion annualized revenue by early 2026, *Claude Code* at $2.5 billion annualized run rate, and **GitHub** Copilot approaching $1 billion.

4.1 Releases

Claude Opus 4.7 became the default coding model in *Claude Code* on April 16. The release added an xhigh effort level between high and max, giving finer reasoning-depth control without the full latency cost of max. The model also introduced explicit task budgets in beta and a /ultrareview command in *Claude Code*. Adaptive thinking replaced extended thinking budgets — setting budget_tokens on the *API* now returns a 400 error.

GPT-5.5 powered Codex from April 23 with a 400,000-*token context window* for paid plans. **OpenAI** granted Pro users 2x Codex usage allowance through May 31. The model is positioned explicitly as an *agent* runtime: the company's announcement framed it as suitable for "messy, multi-part" tasks that require planning, *tool use*, self-checking, and persistence rather than single-turn chat completion.

Cognition completed the acquisition of **Windsurf** for $250 million during April, following **Google**'s acqui-hire of the **Windsurf** founders for $2.4 billion in July 2025. The transaction integrates **Windsurf**'s IDE and Cascade *agent* into **Cognition**'s Devin product line. **Sourcegraph** spun out Amp as a separate company.

4.2 Pricing and licensing

Cursor maintained its $20 per month Pro tier. Devin pricing remained at $20 Core plus $2.25 per *Agent* Compute Unit, the structure introduced in late 2025 to broaden access from the original $500 monthly enterprise tier. **GitHub** Copilot rolled out Opus 4.7 with a 7.5x premium multiplier through April 30 as promotional access; the post-promotion multiplier had not been published at the close of the month. *GPT*-5.5 rolled out across Copilot Pro+, Business, and Enterprise alongside the **OpenAI** Codex integration.

4.3 Adoption signals

JetBrains published the second wave of its AI Pulse developer survey, with January 2026 data drawn from more than ten thousand professional developers. Among specialised AI coding tools, **GitHub** Copilot retained the highest awareness and adoption — 76% awareness, 29% adoption at work — but its growth has stalled since 2025. **Cursor** sits at 69% awareness with adoption growth slowed. *Claude Code* grew from 3% adoption at work in mid-2025 to 18% in January 2026, reaching 24% in the United States and Canada, with the highest customer satisfaction (91%) and net promoter score (54 on a -100 to +100 scale) in the surveyed cohort. The same survey reported that 74% of professional developers worldwide had adopted at least one specialised AI coding tool by January 2026.

5 Industry Trends

Q1 2026 venture and exit data closed during April. Capital concentration in AI reached levels without precedent in the venture record, and the structural constraints on continued AI deployment — power, custom silicon, regulatory readiness — came into clearer view as the *EU AI Act* enforcement deadline approached.

5.1 Funding and valuations

Q1 2026 closed with the highest quarterly venture capital investment ever recorded. KPMG measured $330.9 billion globally; Crunchbase logged $297 billion concentrated in approximately 6,000 companies; CB Insights reported $285.5 billion. The variance reflects different methodologies and reporting cutoffs. AI-related companies absorbed approximately 81% of total deployment.

Four megarounds dominated the quarter: **OpenAI** raised $122 billion at a valuation that placed the company higher than all but a handful of large-cap public technology companies; **Anthropic** raised $30 billion in a Series G led by GIC and Coatue at a $380 billion post-money valuation; **xAI** raised $20 billion before its acquisition by **SpaceX**; and **Waymo** raised $16 billion. **Databricks** closed a $7 billion round; **Polymarket** $2.6 billion; **Shield AI** $2.3 billion. The four leading rounds collectively totalled $188 billion, exceeding all of 2024's global venture deployment.

Two AI foundation labs based in mainland China entered public markets in Q1 via Hong Kong: Z.ai (**Zhipu AI**) and **MiniMax**, each valued above $6 billion at listing. The Hong Kong Stock Exchange has emerged as a viable IPO venue for Chinese AI companies absent direct US listings. **Anthropic** was reported in late April to be running at a $30 billion annualized revenue rate, with secondary investor offers reportedly arriving at valuations approaching $800 billion and early IPO discussions under way. **OpenAI** exceeded $25 billion in annualized revenue.

5.2 Regulation and legal

The European Commission issued €63.2 million in support of AI innovation in health and online safety on April 21. The 2 August 2026 enforcement date for the bulk of *EU AI Act* obligations remained in force at the close of the month, although the *Digital Omnibus on AI* — adopted by the Commission in November 2025 — continued to be negotiated by the Parliament and Council. The *Digital Omnibus* proposes a maximum postponement to 2 December 2027 for certain high-risk system obligations under Annex III. Article 50 transparency requirements for AI-generated content remained on the original 2 August 2026 schedule under all current proposals.

State-level regulation in the United States continued to fill the gap left by the absence of a federal framework. The

Texas Responsible Artificial Intelligence Governance Act took effect on January 1, 2026; the Utah Artificial Intelligence Policy Act applies disclosure requirements to deployers of *generative AI* in regulated and consumer transactions; the Colorado *AI Act* becomes applicable in June 2026. New York City and the federal Equal Employment Opportunity Commission moved on enforcement actions related to AI-driven hiring tools.

The principal *generative AI* copyright cases — New York Times v. **OpenAI** in the Southern District of New York, and Getty Images v. **Stability AI** in the United Kingdom and the United States — entered decisive phases in April. Court rulings on whether *training* on copyrighted data constitutes fair use are expected during Q2 2026.

5.3 Mergers, acquisitions, exits

SpaceX completed the acquisition of **xAI** on terms that valued the combined entity at $1.25 trillion and the **xAI** business itself at $250 billion. The transaction is the largest M&A in corporate history by absolute value at signing. The combined entity owns **Tesla**'s autonomous-driving programmes, the Grok model family, the X social platform, and the Starlink satellite network.

OpenAI continued an active acquisition programme; the company completed six acquisitions during 2026 by mid-April, equalling its full-year acquisition count for 2025. **Anthropic** completed one publicly disclosed acquisition in 2026 — **Vercept**, a software development startup founded in 2024 — adding to the 2025 acquisitions of **Humanloop** and Bun.

Total venture-backed M&A in Q1 reached $56.6 billion per Crunchbase, the third-highest quarterly total since the 2022 downturn. Total Q1 exit value crossed $413.5 billion globally per KPMG, the highest level since Q4 2021.

5.4 Infrastructure

Compute infrastructure constraints emerged as a binding capacity for AI deployment. Projections for the United States indicate a 9 to 18 gigawatt electricity shortfall by 2027 driven by AI data centre construction. **Microsoft**'s announced restart of the Three Mile Island nuclear facility, originally signalled in 2024, advanced through April. **Amazon** continued exploration of dedicated nuclear sites for AWS data centre power.

Custom silicon partnerships diversified during the quarter. **Google** extended its *TPU* partnership with Broadcom; **Anthropic** signed a multi-year compute commitment with **CoreWeave**; Intel announced a collaboration with **Google** on next-generation AI accelerators. **AMD** continued GPU shipments at scale and increased its participation in AI venture rounds via **AMD** Ventures, including the **Cohere** round and the **World Labs** financing.

A research result attributed to a method called TurboQuant — applying 3-bit *quantization* to *KV cache* memory without measurable accuracy loss — was

reported during the quarter, with claims of up to eight-fold speedup in attention computation and at least six-fold reduction in memory usage. Independent verification of the published method was pending at the close of April. Arista Networks raised its 2026 revenue outlook to $11.25 billion, citing AI cluster deployment as the principal driver.

6 Monthly Recap

Chronological summary of the month’s reportable events. Sources are linked in the corresponding section above.

Date	Category	Event	Source / Impact
April 2	LLM	Google releases the Gemma 4 family, four variants under Apache 2.0	Open-weight benchmark leader on Arena AI
April 2	Generative media	Google Vids adds Veo 3.1 video and Lyria 3 / Pro music generation	Ten free Veo generations per month
April 5	LLM	Meta releases Llama 4 Scout and Maverick, open-weight Mixture of Experts	10M-token context on Scout
April 7	LLM	Anthropic confirms Claude Mythos and withholds it from general availability	First ASL-4 trigger in commercial AI
April 8	LLM	Meta releases Muse Spark, the company’s first proprietary model	Reverses Meta’s open-weights posture
April 9	Pricing	OpenAI introduces a $100 ChatGPT Pro tier between Plus and Pro $200	First subdivision of the Pro tier
April 16	LLM	Anthropic releases Claude Opus 4.7	87.6% SWE-bench Verified, 64.3% Pro

Date	Category	Event	Source / Impact
April 20–24	LLM	DeepSeek V4 (1.6T parameters) and xAI Grok 4.3 reach general availability	V4 the largest publicly announced model
April 21	Generative media	OpenAI launches ChatGPT Images 2.0	Replaces GPT Image 1.5; adds Thinking Mode
April 21	Regulation	European Commission allocates €63.2M for AI in health and online safety	Pre-deadline support measure
April 23	LLM	OpenAI releases GPT-5.5 (codename Spud)	First fully retrained base since GPT-4.5
April 24	LLM	GPT-5.5 API access opens	$5 / $30 per million tokens, doubling GPT-5.4
April	LLM	Moonshot Kimi K2.6, Alibaba Qwen 3.6-Plus, Zhipu GLM-5.1, Arcee Trinity 400B released	Open-weight ecosystem expansion
April	M&A	Cognition completes acquisition of Windsurf for $250M	Cascade agent integrated into Devin product line

7 Outlook

GPT-6 timing remains uncertain. Prediction markets had assigned April 2026 some probability of release; that has now slipped. Sam Altman’s most recent on-record statement was "a few more weeks." Whether the next **OpenAI** release continues the *GPT*-5.x line or marks a generational increment to *GPT*-6 has not been resolved.

xAI has indicated Grok 5 for the second quarter, with reported architectural details pointing to a 6-trillion-parameter *Mixture of Experts*. The validity of these reports is partial; treat the parameter count as unconfirmed.

DeepSeek's R2 successor has been reported as delayed by performance shortfalls and chip-export-control constraints affecting access to high-end **NVIDIA** hardware. A Q2 release window appears unlikely on current information.

Claude Opus 4.7 may itself yield to a successor before September. **Anthropic**'s roughly two-month cadence — Opus 4.5 in November 2025, Opus 4.6 in February 2026, Opus 4.7 in April 2026 — implies an Opus 4.8 release in the early summer, with possible introduction of a Mythos-class model in restricted general availability if internal safety reviews complete.

The *Digital Omnibus on AI* is expected to reach political agreement in the Council before June. Whether the proposed postponement of certain high-risk obligations to 2027 takes legal effect before the original 2 August deadline will determine the compliance posture for several thousand companies operating in or selling into the European market.

The first rulings in the principal generative-AI copyright cases — New York Times v. **OpenAI** and Getty Images v. **Stability AI** — are expected during Q2.