AI News

Trends — March 2026

2026-05-02 - 14:22

Compiled with Claude AI

1 Executive Summary

March 2026 produced a denser model release cadence than any previous month on record. Nine frontier or near-frontier text models reached availability or preview. **OpenAI** shipped *GPT*-5.4 with native *computer use* on the 5th, the first general-purpose **OpenAI** model with desktop control built into the base. **Google** released Gemini 3.1 Flash-Lite on the 3rd. **Mistral** AI shipped six products in fifteen days, including the consolidated **Mistral** Small 4 *reasoning model* and the Voxtral text-to-speech system. *Open-weight* ecosystem activity intensified with **Alibaba**'s Qwen 3.5 series, **Xiaomi**'s MiMo-V2-Pro, **MiniMax**-M2.7, and **NVIDIA**'s Nemotron 3 Super and VoiceChat all reaching availability.

OpenAI announced on March 24 that the Sora app and *API* would be discontinued, with the consumer app shutting down on April 26 and the *API* on September 24. The company cited compute shortages and cost pressures. The decision marks the first sunset of a major AI consumer product brand by a frontier lab. On March 26, an **Anthropic** data store misconfiguration briefly exposed approximately three thousand internal files, including a draft blog post describing a model internally codenamed Capybara that scored above any publicly released system on coding and science benchmarks.

The Q1 funding picture closed with **OpenAI**'s record $122 billion round confirmed final on March 31 — the largest venture financing in history. The *Digital Omnibus on AI* reached political agreement on March 11, proposing a maximum postponement to 2 December 2027 for certain high-risk system obligations under the *EU AI Act*. Article 50 transparency obligations remained on the original 2 August 2026 schedule.

2 Large Language Models

Nine text models reached general availability or preview during March, the highest single-month total on independent trackers. The release cadence reflects compressed competition between four labs operating on overlapping schedules — **OpenAI**, **Google**, **Mistral**, and the broad Chinese ecosystem.

2.1 Releases and updates

Google released Gemini 3.1 Flash-Lite on March 3 in preview through the Gemini *API*, **Google** AI Studio, and Vertex AI. The model targets the cost-sensitive tier with $0.25 per million input *tokens* and $1.50 per million output *tokens* — half the price of Gemini 3 Flash. Native multimodality across text, image, audio, and video is preserved, alongside a 1-million-*token context window* inherited from the broader Gemini 3 family. Gemini 3.1 Flash-Lite is based on Gemini 3 Pro and trained on **Google**'s *TPUs*.

OpenAI released *GPT*-5.4 on March 5, simultaneously across ChatGPT (as *GPT*-5.4 Thinking), the *API*, and Codex. The release is the first general-purpose **OpenAI** model with native computer-use capabilities built into the base rather than offered as a separate Codex variant. *GPT*-5.4 absorbs the

prior GPT-5.3-Codex line into the mainline, simplifying model selection. The headline benchmark is OSWorld-Verified at 75%, surpassing the 72.4% human-expert baseline — the first model to do so. Twelve days later, on March 17, **OpenAI** released GPT-5.4 mini and GPT-5.4 nano: mini sits at $0.75 per million input and $4.50 per million output, offered to free-tier ChatGPT users; nano is API-only.

Mistral shipped six products between March 16 and March 31. **Mistral** Small 4, released March 16, merges three previously separate product lines — Magistral (reasoning), Pixtral (multimodal vision), and Devstral (agentic coding) — into a single model with configurable reasoning effort. The model is a 119-billion-parameter Mixture of Experts with 6.5 billion active parameters per forward pass, available under Apache 2.0 at $0.15 per million input tokens. On the same day, **Mistral** released Leanstral, a formal-proof agent for Lean 4 that beats Claude Sonnet 4.6 by 8 points at pass@16 on FLTEval at fifteen times lower cost. Voxtral, a text-to-speech system released March 23, runs on a single GPU at 8 GB BF16 weights. **Mistral** Forge, an enterprise training platform, was announced at **NVIDIA** GTC on March 17.

xAI promoted Grok 4.20 Beta 2 to general availability on March 31, completing the four-agent multi-agent architecture: the Grok coordinator and three specialised sub-agents (research, logic, contrarian analysis) running in parallel and cross-verifying outputs. Grok 4.20 Beta posted a 22% hallucination rate, reportedly the lowest measured across publicly available frontier models in independent evaluation.

Other March releases included **Alibaba**'s Qwen 3.5 series — eight variants from 0.8 billion to 397 billion parameters, the largest comparable open-weight family released in a single month — **Xiaomi**'s MiMo-V2-Pro at one trillion parameters, **MiniMax**-M2.7, and **NVIDIA**'s Nemotron 3 Super and Nemotron 3 VoiceChat models.

2.2 Benchmarks and capabilities

Computer use crossed the human baseline for the first time. GPT-5.4's 75% on OSWorld-Verified exceeded the 72.4% reported for human experts on the same benchmark. Claude Opus 4.6 had previously approached but not crossed this threshold.

The Artificial Analysis Intelligence Index, a composite reasoning score, finished March with three models effectively tied at the top: Gemini 3.1 Pro Preview at 57.18, GPT-5.4 (xhigh) at 57.17, and GPT-5.3 Codex (xhigh) at 54. Claude Opus 4.6 with adaptive reasoning sat at 53. The narrowness of these gaps — under 0.1 points between the top two — indicates that the composite reasoning benchmark has effectively saturated for the publicly available frontier.

Open-weight competitiveness improved further. GLM-5.1 from **Zhipu AI** ranked first among *open-weight* models on the *Intelligence Index* at 51 points. The gap between best *open-weight* and best proprietary narrowed to approximately six index points — down from twelve at the start of the quarter.

OpenAI's *GPT*-5.4 also reported a 33% reduction in factual errors compared to *GPT*-5.2 on internal benchmarks, alongside a tool-search architecture intended to reduce *token* consumption in tool-heavy workflows by up to 47%. Independent reproduction of the tool-search efficiency claim was pending at the close of the month.

2.3 Pricing and access

GPT-5.4 entered at $2.50 per million input and $15 per million output *tokens* for the standard tier, with *GPT*-5.4 Pro at $30 per million input and $180 per million output. The pricing represents a notable undercut against *Claude* Opus 4.6, which sat at $5 input and $25 standard / $75 output Pro. *GPT*-5.4 mini and nano, released March 17, fall at $0.75 / $4.50 for mini and lower for nano (*API* only).

Gemini 3.1 Flash-Lite at $0.25 input and $1.50 output is approximately half the price of Gemini 3 Flash and one-tenth the price of *GPT*-5.4 standard, while preserving the 1-million-*token* context. **Mistral** Small 4 at $0.15 per million input *tokens* is positioned as one of the cheapest *multimodal* reasoning models on offer — five times cheaper than *GPT*-5.4 mini on input. The model is also fully *open-weight* under Apache 2.0, allowing self-hosting at zero licence cost.

Mistral hit $400 million in annualized recurring revenue in January 2026, up from approximately $20 million a year earlier per CEO Arthur Mensch's public statement, against a $13.8 billion valuation reached during the quarter.

3 Generative Media

March produced one of the most consequential events of the quarter for generative media — the announced shutdown of the Sora consumer product line by **OpenAI** — alongside continued capability progress in video and audio. The image segment remained relatively quiet on standalone releases.

3.1 Image

No major flagship image model was released by **Black Forest Labs**, **Stability AI**, or Midjourney during March. The category has matured: capability gains continue inside the editor-integrated workflows of **Adobe** Firefly, ChatGPT, and **Google** Vids rather than through standalone *text-to-image* tool releases. A research preview titled ImageCritic, released during the month, addresses fine-grained inconsistencies in AI-generated images by detecting reference image mismatches against a target.

3.2 Video

On March 24, **OpenAI** announced via X that the Sora consumer app and the Sora *API* would be discontinued. The app was scheduled for shutdown on April 26 and the *API* on September 24. **OpenAI** did not provide a specific reason in the shutdown notice; subsequent reporting linked the decision to compute shortages, cost pressures, and a strategic reallocation toward core enterprise products. The **OpenAI** partnership with **Disney** that licensed certain characters for use within Sora was also winding down. The shutdown marks the first sunset of a major AI consumer product brand by a frontier lab.

The market continued with **Runway** Gen-4.5, Kling 3.0 from **Kuaishou**, Veo 3.1 from **Google**, Pika 2.0, **ByteDance** Seedance 2.0, and Luma Ray 3. Kling 3.0, released in February, introduced multi-shot sequences with subject consistency across camera angles. Seedance 2.0, also released in February, offered unified audio-visual joint generation with phoneme-level lip-sync in eight languages. Both models reached broader *API* distribution during March via FAL.AI and partner platforms.

3.3 Music and audio

Suno released version 5.5 in March, adding custom voice cloning, personalized model *training*, and studio-quality tracks of eight minutes or longer. **Suno** reported 2 million paid subscribers, $300 million in annual revenue, and a $2.45 billion valuation by mid-quarter.

Mistral Voxtral TTS, released March 23, is the company's first audio model — an *open-weight* system (CC BY-NC 4.0) built on Ministral 3B. The default BF16 weights run on a single GPU with 16 GB of VRAM; *quantized* weights run on edge devices. The release positions **Mistral** as a direct competitor to **ElevenLabs** in the text-to-speech segment. **Udio** remained constrained by the October 2025 Universal Music Group settlement, which limited paid users to streaming-only access without download capability.

4 Coding and Developer Tools

The coding tools segment continued growth at a pace that compresses each quarter's revenue benchmarks against the previous one. **Cursor** reached $2 billion annualized run rate; *Claude Code* reached $2.5 billion annualized in February; **GitHub** Copilot approached $1 billion. March added new product capability primarily through the unification of **OpenAI**'s Codex into the *GPT*-5.4 mainline and through **Replit**'s parallel-task *agent*.

4.1 Releases

Replit released *Agent* 4 on March 11, introducing parallel task forking that automatically resolves merge conflicts in approximately 90% of cases. The release coincided with a $400 million Series D round at a $9 billion valuation. **Replit** *Agent* 4 is included in the standard **Replit** subscription rather than offered as a premium add-on.

GPT-5.4's absorption of *GPT*-5.3-Codex's coding capabilities into the mainline, on March 5, eliminated the prior need to choose between a coding-specialised model and a general model. *GPT*-5.4 entered Codex as the primary model. *GPT*-5.3-Codex was not deprecated; **OpenAI** continued support across both lines. *GPT*-5.4 mini, released March 17, became available to free-tier ChatGPT users with *GPT*-5.4 nano available through the **OpenAI** *API* only — the first time free-tier users received access to a frontier-class coding model from **OpenAI**.

Mistral Forge, announced at **NVIDIA** GTC on March 17, allows enterprises to train frontier-grade models on proprietary data. The service targets enterprise customers building domain-specific models without rebuilding from scratch. **Mistral** Leanstral, released March 16, addresses a narrower use case — formal proof engineering in Lean 4 — and reports beating *Claude* Opus 4.6 on FLTEval at one ninety-second of the cost.

4.2 Pricing and licensing

GPT-5.4's pricing at $2.50 / $15 per million *tokens* at the standard tier represented a significant undercut against *Claude* Opus 4.6 within the developer tools segment. *GPT*-5.4 Pro at $30 / $180 sits roughly 2.4x cheaper than *Claude* Opus 4.6 on output costs while delivering competitive coding performance.

Replit's $400 million Series D placed it at a $9 billion valuation, with *Agent* 4 priced into the standard **Replit** subscription rather than as a premium tier. **Mistral** Small 4 at $0.15 per million input and $0.60 per million output stands out as the cheapest *multimodal reasoning model* in the segment for self-hosted deployment under Apache 2.0.

4.3 Adoption signals

The dominant adoption signal from March was the broadening of free-tier access. *GPT*-5.4 mini in the free ChatGPT tier, combined with the cost reduction across Gemini 3.1 Flash-Lite and **Mistral** Small 4, expanded the working population of *LLM*-assisted developers. **Cursor** and *Claude Code* remained the dominant combination at the high end, with combined annualized revenue exceeding $4.5 billion across the two products. Devin, by **Cognition**, reportedly approached $150 million annualized run rate by the end of Q1.

5 Industry Trends

Q1 venture capital deployment closed at the largest quarterly total ever recorded. Concurrently, the EU regulatory framework moved closer to enforcement, with the *Digital Omnibus* political agreement on March 11 reshaping the pre-deadline compliance landscape. Compute infrastructure constraints — first surfaced as a financial concern in 2025 — became a binding operational constraint in the form of **OpenAI**'s Sora discontinuation.

5.1 Funding and valuations

OpenAI's record $122 billion round was confirmed final on March 31, after a tranche-by-tranche disclosure throughout the quarter. Major participants included Andreessen Horowitz, D.E. Shaw, MGX, TPG, and T. Rowe Price, alongside earlier investors. The round is the largest venture financing in history.

Anthropic's $30 billion Series G, led by GIC and Coatue, valued the company at $380 billion post-money. Other significant Q1 closings included **xAI** at $20 billion, **Waymo** at $16 billion, **Databricks** at $7 billion, **Polymarket** at $2.6 billion, and **Shield AI** at $2.3 billion. **Mistral** hit $400 million in annualized recurring revenue and a $13.8 billion valuation. Two AI foundation labs based in mainland China entered public markets via the Hong Kong Stock Exchange during the quarter: Z.ai (**Zhipu AI**) and **MiniMax**, each valued above $6 billion at listing.

5.2 Regulation and legal

The *Digital Omnibus on AI* reached political agreement on March 11. The proposal extends certain high-risk system obligations under the *EU AI Act* to 2 December 2027 — a postponement of approximately sixteen months over the original 2 August 2026 schedule. Article 50 transparency obligations for AI-generated content remain on the original schedule. The European Parliament and the Council of the EU continued formal negotiation through the end of March; a formal vote was expected during April or May.

The European Commission published the second draft of the Code of Practice on Marking and Labelling of AI-generated Content on March 5. The code is voluntary but functions as a compliance reference for Article 50 obligations. It establishes recommended technical approaches to watermarking, metadata labelling, and disclosure for synthetic content.

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 — continued through pre-trial motions during March. Initial summary judgement rulings were expected during Q2.

5.3 Mergers, acquisitions, exits

OpenAI completed six acquisitions during Q1 alone, equalling its full-year acquisition count for 2025. **Anthropic** completed one publicly disclosed acquisition during the quarter — **Vercept**, a software development startup founded in 2024 — adding to its 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. The largest planned transactions of the quarter included Capital One's $5.15 billion agreement to acquire Brex and Savvy Games Group's $6 billion planned acquisition of **ByteDance**'s gaming platform Moonton.

5.4 Infrastructure

NVIDIA GTC took place from March 17 to 21 in San Jose. The conference featured announcements of Nemotron 3 Super, Nemotron 3 VoiceChat, and the **MISTRAL** Forge enterprise platform. **NVIDIA**'s Cosmos and GR00T open models for robotics also progressed during the quarter. **AMD** continued GPU shipments at scale and increased its participation in AI venture rounds via **AMD** Ventures.

The **ANTHROPIC** data store misconfiguration disclosed on March 26 exposed approximately three thousand internal files briefly. The exposed material included a draft blog post describing a model internally codenamed Capybara, which had reportedly scored 93.9% on *SWE-bench* Verified and 94.6% on *GPQA Diamond* — both above any publicly available system at the close of March. **ANTHROPIC** confirmed the exposure but did not, by the end of the month, formally announce the model.

The Sora shutdown announcement on March 24 was the principal infrastructure-related event for **OPENAI**. The discontinuation reflected the binding nature of compute shortages on AI product portfolios; **OPENAI** was implicitly choosing to reallocate compute capacity from consumer media generation to enterprise coding and reasoning workloads. The strategic message — that compute is now scarce enough to force product portfolio decisions at frontier labs — was new in March.

6 Monthly Recap

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

Date	Category	Event	Source / Impact
March 3	LLM	Google releases Gemini 3.1 Flash-Lite in preview	$0.25 / $1.50 per million tokens
March 5	LLM	OpenAI releases GPT-5.4 with native computer use	First model to exceed human OSWorld baseline
March 5	Regulation	European Commission publishes second draft of AI Content Marking Code of Practice	Reference for Article 50 compliance

Date	Category	Event	Source / Impact
March 11	Coding	Replit releases Agent 4 with parallel task forking	$400M Series D at $9B valuation
March 11	Regulation	Digital Omnibus on AI reaches political agreement in Brussels	Proposes postponement to 2 December 2027
March 16	LLM	Mistral Small 4 released, merges Magistral, Pixtral, Devstral	$0.15 / $0.60 input / output, Apache 2.0
March 16	Coding	Mistral Leanstral released, formal proof agent for Lean 4	Beats Sonnet 4.6 at pass@16, 15x cheaper
March 17	LLM	OpenAI releases GPT-5.4 mini and nano	Mini available on free ChatGPT tier
March 17–21	Industry	NVIDIA GTC conference in San Jose	Nemotron 3 Super, VoiceChat releases
March 23	Generative media	Mistral Voxtral TTS released	Open-weight, 8 GB single-GPU footprint
March 24	Generative media	OpenAI announces Sora app and API discontinuation	App ends 26 April; API ends 24 September
March 26	LLM	Anthropic data store misconfiguration exposes draft of leaked model	Codenamed Capybara, top benchmark scores

Date	Category	Event	Source / Impact
March 31	LLM	xAI promotes Grok 4.20 Beta 2 to full release	Multi-agent architecture, 22% hallucination rate
March 31	Funding	OpenAI $122 billion round confirmed final	Largest venture financing in history
March	LLM	Alibaba Qwen 3.5 series, Xiaomi MiMo-V2-Pro, MiniMax-M2.7 released	Open-weight ecosystem expansion
March	Generative media	Suno 5.5 released	Custom voice cloning, eight-minute tracks

7 Outlook

OpenAI's release cadence — six weeks between *GPT*-5.2 and *GPT*-5.3 Codex, six weeks between *GPT*-5.3 and *GPT*-5.4 — suggests another release in mid-to-late April. The naming may continue as *GPT*-5.5 or jump to a new generation. Sam Altman's recent comments have not committed to a specific window.

Anthropic's leaked draft from March 26 describes a model internally codenamed Capybara with *benchmark* scores above any publicly released system. The company's response — formal release, restricted access, or quiet retirement — will define the safety-versus-capability axis for the rest of the quarter and is expected during April.

Anthropic's roughly two-month cadence implies a *Claude* Opus 4.7 release in the early second quarter, following Opus 4.6 in February. The release is likely to arrive before the end of April.

The **Replit** *Agent* 4 release establishes a pattern likely to be followed by other coding agents during April: parallel autonomous task execution rather than serial planning. **Cursor** and *Claude Code* background agents already operate similarly; the differentiator in Q2 is likely to be the merge conflict resolution rate and the integration depth with code review systems.

The *Digital Omnibus* political agreement is expected to receive a formal vote in April or May. The legal effect of any postponement of high-risk obligations under the *EU AI Act* remains contingent on this vote and on subsequent transposition by

member states. Companies operating in the EU should treat the 2 August 2026 deadline as binding until the formal postponement is enacted.

The shutdown of the Sora app on April 26 will redirect **OpenAI**'s compute capacity. Whether that capacity is redeployed to a successor consumer media product or to enterprise coding and reasoning workloads will be a notable indicator of strategic priority.

The New York Times v. **OpenAI** and Getty Images v. **Stability AI** cases continue toward decisive phases. Initial summary judgement rulings are expected during Q2, with the New York Times case likely to rule first given its more advanced procedural position.