AI News

Trends — February 2026

2026-05-02 - 14:22

Compiled with Claude AI

1 Executive Summary

February 2026 produced the most concentrated *frontier model* release window observed to date. **Anthropic** shipped *Claude* Opus 4.6 on the 5th and *Claude* Sonnet 4.6 on the 17th. **Google** released Gemini 3.1 Pro on the 19th. **xAI** released Grok Imagine 1.0 on the 2nd, **OpenAI** shipped *GPT*-5.3-Codex on the 5th, and **Zhipu AI** released GLM-5 on the 11th. Across the month, more than ten *open-weight* or proprietary frontier-class models reached availability — a pace independent observers characterized as the densest model release window in AI history.

The 1-million-*token context window* became the standard for frontier models. *Claude* Opus 4.6 and Sonnet 4.6 both shipped with 1M context at standard pricing without beta headers. **Anthropic**'s Opus 4.6 set the highest task-completion time horizon estimated by *METR* — 14 hours and 30 minutes at the 50% mark — and the highest reported *MRCR v2* score at 1M *tokens* (78.3%). Gemini 3.1 Pro doubled its predecessor's *ARC-AGI-2* score (31.1% to 77.1%) and reached 94.3% on *GPQA Diamond*, the highest score reported on that *benchmark* at the time.

Anthropic raised $30 billion in a Series G funding round on February 12 at a $380 billion post-money valuation — the second-largest private financing round in technology history, behind only **OpenAI**'s $40 billion raise of 2025. The round was led by GIC and Coatue, with co-leads from D. E. Shaw Ventures, Dragoneer, Founders Fund, ICONIQ, and MGX. **Anthropic** disclosed run-rate revenue of $14 billion at the time of the round, with *Claude Code* at $2.5 billion annualized and weekly active users having doubled since January.

To watch in March: **OpenAI**'s expected release cadence step (*GPT*-5.4 timeline), the second draft of the EU Code of Practice on Marking and Labelling of AI-generated Content, Q1 venture deployment closing data, and continued *open-weight* competition from **Alibaba**, Zhipu, and **DeepSeek**.

2 Large Language Models

February's release cadence was tighter than any month previously recorded. The release windows of **OpenAI**, **Anthropic**, **Google**, **xAI**, **Alibaba**, Zhipu, and **ByteDance** overlapped within hours of each other on multiple occasions.

2.1 Releases and updates

Anthropic released *Claude* Opus 4.6 on February 5, available immediately through *claude*.ai, the *Claude API*, **Amazon** Bedrock, **Google** Cloud Vertex AI, and **Microsoft** Foundry on Azure. Opus 4.6 ships with a 1-million-*token* input context and up to 128K output *tokens*. The model introduced *agent teams* — multiple *Claude* instances running in parallel with peer-to-peer communication via what **Anthropic** called the Mailbox Protocol — and context compaction, which automatically summarizes earlier portions of a conversation as the *context window* approaches its limit. *METR*'s task-completion time horizon estimate places Opus 4.6 at 14 hours and 30 minutes at the 50% mark and 1 hour 3 minutes at the 80% mark, the highest

reported among frontier models. *Long-context* retrieval on *MRCR v2* at 1M *tokens* reached 78.3%, ahead of *GPT*-5.4 at 36.6% and Gemini 3.1 Pro at 18.3%.

Anthropic followed with *Claude* Sonnet 4.6 on February 17. Pricing of $3 per million input *tokens* and $15 per million output *tokens*, with 1-million-*token* context at standard pricing, made it the lowest-priced model with that context length on the market at the time. **Anthropic** reported that early-access users preferred Sonnet 4.6 to Sonnet 4.5 in approximately 70% of head-to-head tests, and to the prior flagship Opus 4.5 in 59% of tests. Sonnet 4.6 leads on *GDPval-AA*, an evaluation of economically valuable professional tasks, with an Elo score of 1633 against Opus 4.6 at 1606 and Gemini 3.1 Pro at 1317.

Google released Gemini 3.1 Pro on February 19 in preview through the Gemini *API*, **Google** AI Studio, Vertex AI, Gemini CLI, Android Studio, and the Gemini app for Pro and Ultra subscribers. Gemini 3.1 Pro is the first version-point increment in the Gemini 3 series and integrates the reasoning techniques first deployed in Gemini 3 Deep Think. The model scored 77.1% on *ARC-AGI-2* — more than double the 31.1% recorded by Gemini 3 Pro three months earlier — and 94.3% on *GPQA Diamond*, the highest score reported on that *benchmark* at the time of release. Pricing came in at $2 per million input *tokens* and $12 per million output *tokens*, undercutting both *Claude* Opus 4.6 and *GPT*-5.3-Codex. The model supports a 1-million-*token context window* with *multimodal* input across text, image, audio, and video.

OpenAI shipped *GPT*-5.3-Codex on February 5, the first **OpenAI** release dedicated to coding workflows under the Codex line. *GPT*-5.3-Codex remained available alongside *GPT*-5.2 Instant and the general *GPT*-5.3 variants. The model targeted *SWE-bench*-class evaluations and reached the upper tier of leaderboards on terminal coding tasks. Gemini 3 Deep Think, released the week of February 12, was used by **Google** to solve problems in science, research, and engineering, including the disproof of a decade-old mathematics conjecture.

xAI released Grok Imagine 1.0 on February 2, providing a publicly documented *multimodal* generation *API* for both image and video output. **xAI** followed with Grok 4.2 Beta on approximately February 17, ahead of the full Grok 4.20 Beta 2 release in March. Grok 4.2 Beta retained the four-*agent* multi-*agent* architecture with the Grok coordinator and three specialised sub-agents.

February also produced unusually intense *open-weight* ecosystem activity. **Zhipu AI** released GLM-5 on February 11, a 744-billion-parameter *Mixture of Experts* model. **ByteDance** shipped Kimi K2.5 with reasoning capabilities and Doubao 2.0. **MiniMax** launched M2.5 and M2.5 Lightning on February 12, posting strong *SWE-bench* scores at low cost as open models.

DeepSeek released V3.2. **Alibaba**'s Qwen 3.5 family began rolling out mid-month with an *agent*-focused architecture.

2.2 Benchmarks and capabilities

Three benchmarks moved meaningfully during February. *ARC-AGI-2* — designed to test novel-pattern recognition that resists *training*-set memorization — doubled at the frontier from Gemini 3 Pro's 31.1% to Gemini 3.1 Pro's 77.1%. *GPQA Diamond*, a graduate-level science test, reached 94.3% on Gemini 3.1 Pro — the highest score reported on the *benchmark* to that date. The 1-million-*token* context retrieval *benchmark MRCR v2* reached 78.3% on *Claude* Opus 4.6 at 1M *tokens*, against 36.6% on *GPT*-5.4 and 18.3% on Gemini 3.1 Pro at the same context length.

The LMSys *Chatbot* Arena, which reflects crowd-sourced user preferences across diverse tasks, registered Gemini 3.1 Pro Preview at 1500 Elo and *Claude* Opus 4.6 variants at 1504 Elo — effectively tied at the top during the month.

Long task-completion horizons emerged as a meaningful axis of differentiation. *METR*'s estimate places Opus 4.6 at 14 hours and 30 minutes for 50% completion. The metric matters because the practical limit of *agentic* deployment is task length — a model that loses coherence after two hours cannot complete an eight-hour analytical workflow no matter how high its single-step *benchmark* scores.

Computer use approached the human-expert baseline. Sonnet 4.6's improvement on *OSWorld-Verified*, combined with its *prompt*-injection resistance now on par with Opus 4.6, made browser-based and desktop-based *agentic* deployment more practical at lower cost than the Opus tier required.

2.3 Pricing and access

Claude Opus 4.6 entered at $5 per million input *tokens* and $25 per million output *tokens*, with up to 128K output *tokens* per request. **Anthropic** introduced a Fast Mode for Opus 4.6 delivering up to 2.5x faster output at premium pricing of $30 per million input and $150 per million output — same intelligence, faster turnaround.

Claude Sonnet 4.6 entered at $3 per million input and $15 per million output *tokens*. Both Opus 4.6 and Sonnet 4.6 included 1-million-*token* context at standard pricing without beta headers. The previously charged premium for requests exceeding 200K *tokens* was retained at the standard rate but the *long-context* surcharge would later be removed entirely on March 13.

Gemini 3.1 Pro at $2 per million input and $12 per million output *tokens* undercut both *Claude* Opus 4.6 and **OpenAI**'s *GPT*-5.x line on per-*token* pricing while delivering competitive *benchmark* scores. *Open-weight* options at near-zero licence cost — Zhipu GLM-5, **MiniMax** M2.5, **ByteDance** Kimi

K2.5, **Alibaba** Qwen 3.5, **DeepSeek** V3.2 — expanded the cost floor downward across the segment.

3 Generative Media

Generative media activity in February concentrated in video, with two new flagship architectures and one new entrant in the space, alongside continued capability advances in image and audio.

3.1 Image

xAI released Grok Imagine 1.0 on February 2, the company's first *multimodal* generation product. The release covered both image and video output through a publicly documented *API* and partner platforms, allowing third-party integration into product pipelines without building a separate media stack. **Adobe**, **Black Forest Labs**, **Stability AI**, and Midjourney did not release new flagship image models during the month.

3.2 Video

Kling 3.0, from **Kuaishou**, was released during February with multi-shot sequence support of three to fifteen seconds and subject consistency across different camera angles — a meaningful technical step that simplified narrative video workflows previously requiring manual frame stitching. Multi-character native audio with voice reference allowed consistent voices across cuts.

ByteDance Seedance 2.0 was released during February as the first AI video model with unified audio-visual joint generation. Generation produced sound and image simultaneously rather than as a post-processed audio overlay. The model supported twelve-file *multimodal* input, phoneme-level lip-sync in eight languages, and identity lock for character consistency across scenes. Both Kling 3.0 and Seedance 2.0 reached broader *API* distribution through FAL.AI and partner platforms.

Sora 2 and Sora 2 Pro continued in service from **OpenAI** — **OpenAI**'s announcement of the Sora discontinuation would not arrive until March 24.

3.3 Music and audio

Suno and **Udio** did not release new flagship music models during February; both continued under their existing version lines (**Suno** 5.x, **Udio** constrained by the October 2025 Universal Music Group settlement). **ElevenLabs** continued integration partnerships with video generation pipelines. **Google**'s ProducerAI, built on the Lyria 3 architecture released in late 2025, expanded inside **Google** Labs as a music generation tool for Workspace.

4 Coding and Developer Tools

February's coding tools activity was driven by the **Anthropic** Series G announcement and the **OpenAI** *GPT*-5.3-Codex release. Both events underscored the segment's emergence as the largest revenue category in commercial AI.

4.1 Releases

Anthropic launched *Claude Code* Security during February, an *agentic* feature that reviews codebases to identify security vulnerabilities. The feature operates inside the existing *Claude Code* product and was made available to existing subscribers. *Claude* Opus 4.6, released February 5, became the default coding model in *Claude Code*; the *agent teams* capability allowed multiple *Claude* instances to work in parallel on different parts of a project.

OpenAI released *GPT*-5.3-Codex on February 5 as a coding-specialised model under the Codex line. The release continued the pattern of separate Codex variants for coding workflows, a structure that *GPT*-5.4 would consolidate one month later.

4.2 Pricing and licensing

Claude Code revenue at $2.5 billion annualized was disclosed during the **Anthropic** Series G announcement on February 12, more than double the level at the start of the year and an indicator of how rapidly coding-tool revenue compounded. *Claude Code* business subscriptions had quadrupled since the start of 2026, and enterprise users represented over half of *Claude Code* revenue. **Cursor** remained the largest standalone IDE-based AI coding product, while **GitHub** Copilot continued at scale.

Devin pricing remained at $20 Core plus $2.25 per *Agent* Compute Unit, the structure introduced in late 2025. The shift from the original $500 monthly enterprise tier broadened access materially.

4.3 Adoption signals

Anthropic's disclosed *Claude Code* metrics during the Series G announcement provided the clearest single set of adoption data points of the quarter. *Claude Code* authored approximately 4% of all **GitHub** public commits worldwide — double the percentage from one month prior. Weekly active users had doubled since January 1. Eight of the Fortune 10 were now *Claude* customers, and customers spending more than $100,000 annually had grown sevenfold over the past year. Customers spending more than $1 million annually exceeded 500, against a dozen two years earlier.

The data point most relevant to the broader category was that 79% of **OpenAI** users also paid for **Anthropic**, per Ramp aggregated payment data. Coding-tool adoption was therefore additive rather than substitutional within the developer market. One in five businesses using Ramp paid for **Anthropic**, against one in twenty-five a year earlier.

5 Industry Trends

February closed the third month of the most concentrated AI capital deployment cycle on record. The **Anthropic** Series G alone matched the average annual venture

deployment of pre-2020 cycles. Concurrently, the *EU AI Act*'s August 2 enforcement deadline drew nearer with the *Digital Omnibus* negotiation continuing in Brussels.

5.1 Funding and valuations

Anthropic's Series G was announced on February 12. The $30 billion raise valued the company at $380 billion post-money, more than doubling the September 2025 Series F valuation of $183 billion. The round was led by GIC and Coatue, with co-leads from D. E. Shaw Ventures, Dragoneer, Founders Fund, ICONIQ, and MGX. Other significant investors included Accel, General Catalyst, Jane Street, **Microsoft**, **Nvidia**, and the Qatar Investment Authority. The round included portions of the previously announced **Microsoft** (up to $5 billion) and **Nvidia** (up to $10 billion) commitments from November 2025.

Anthropic disclosed financial data alongside the round. Run-rate revenue had reached $14 billion, growing more than tenfold annually in each of the prior three years since the company first earned revenue. Dario Amodei told CNBC that approximately 80% of **Anthropic**'s business came from enterprises. The Series G is the second-largest private technology financing round in history, behind **OpenAI**'s $40 billion raise of 2025.

OpenAI itself was reportedly seeking an additional $100 billion in February, in talks that would close at $122 billion at the end of March. 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**, while not closing a round during the quarter, hit $400 million in annualized recurring revenue per CEO Arthur Mensch's public statement.

5.2 Regulation and legal

The *EU AI Act* 2 August 2026 enforcement deadline remained in force at the close of the month. The *Digital Omnibus on AI* — adopted by the European Commission in November 2025 — continued through Parliament and Council negotiation, with a political agreement targeted for March. The *Digital Omnibus* proposed a maximum postponement to 2 December 2027 for certain high-risk system obligations under Annex III, while preserving the original 2 August 2026 schedule for Article 50 transparency obligations.

US state-level regulation continued to fill the federal gap. The *Texas Responsible Artificial Intelligence Governance Act*, which had taken effect on January 1, 2026, entered its first month of practical enforcement. The Utah Artificial Intelligence Policy Act applied disclosure requirements to deployers of *generative AI* in regulated and consumer transactions. The Colorado *AI Act*, due to apply in June, drove pre-compliance preparation across enterprise deployments.

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 February. No decisive rulings were issued during the month.

5.3 Mergers, acquisitions, exits

EPIC GAMES announced on February 22 the acquisition of **MESHCAPADE**, an AI startup specialising in the creation and animation of hyper-realistic digital human models from video recordings. The transaction extended Epic's investment in tooling for AI-generated character animation in real-time game and film engines.

ANTHROPIC's acquisition activity continued during the quarter with the publicly disclosed acquisition of **VERCEPT**, a software development startup founded in 2024. The transaction added to **ANTHROPIC**'s 2025 acquisitions of **HUMANLOOP** and Bun. **OPENAI** continued its Q1 acquisition programme.

Z.ai (**ZHIPU AI**) and **MINIMAX** progressed toward Hong Kong Stock Exchange listings during the quarter. Both companies reached public-market entry valuations above $6 billion, with formal listings completing during Q1.

5.4 Infrastructure

Compute capacity remained the binding constraint on AI deployment. **NVIDIA**, **AMD**, and Intel announced expanded production schedules during the month, alongside continued progress on **GOOGLE**'s *TPU* partnership with Broadcom and **ANTHROPIC**'s multi-year compute commitments.

ANTHROPIC's Series G announcement explicitly cited infrastructure expansion as a primary use of proceeds. CFO Krishna Rao's statement framed the funding as enabling continued enterprise-grade product development at the scale demanded by Fortune-class customers. The pattern was consistent across the major frontier labs: capital deployment in 2026 is driven not principally by *training* compute requirements but by *inference* capacity to serve high-volume *agentic* workloads.

6 Monthly Recap

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

Date	Category	Event	Source / Impact
February 2	Generative media	xAI releases Grok Imagine 1.0 multimodal generation API	First xAI image and video product
February 5	LLM	Anthropic releases Claude Opus 4.6	1M context standard, 14h30 task horizon

Date	Category	Event	Source / Impact
February 5	LLM	OpenAI releases GPT-5.3-Codex	Coding-specialised model under Codex line
February 11	LLM	Zhipu AI releases GLM-5	744B parameter open-weights MoE
February 12	Funding	Anthropic Series G $30 billion at $380 billion post-money	Second-largest VC round in tech history
February 12	LLM	MiniMax releases M2.5 and M2.5 Lightning	Open-weights, strong SWE-bench at low cost
February 12	LLM	Google demonstrates Gemini 3 Deep Think	Disproof of decade-old mathematics conjecture
February 17	LLM	Anthropic releases Claude Sonnet 4.6	$3 / $15 per million tokens, 1M context
February 17–18	LLM	xAI releases Grok 4.2 Beta	Pre-release of Grok 4.20 Beta 2 line
February 19	LLM	Google releases Gemini 3.1 Pro in preview	77.1% ARC-AGI-2, 94.3% GPQA Diamond
February 22	M&A	Epic Games acquires Meshcapade	Hyper-realistic digital human modelling

Date	Category	Event	Source / Impact
February	LLM	ByteDance Kimi K2.5 with reasoning, Doubao 2.0, Alibaba Qwen 3.5, DeepSeek V3.2	Open-weight ecosystem expansion
February	Generative media	Kling 3.0 (Kuaishou) and Seedance 2.0 (ByteDance) reach broader distribution	Multi-shot sequences, unified audio-video
February	Coding	Anthropic launches Claude Code Security	Agentic vulnerability review in Claude Code

7 Outlook

OpenAI's release cadence implies a successor to *GPT*-5.3 in early-to-mid March. The naming may continue as *GPT*-5.4 or jump to a new generation. The model is expected to consolidate the *GPT*-5.3-Codex coding capabilities into the mainline rather than maintaining the separate Codex variant.

Mistral, having committed publicly to a $400 million ARR run-rate at a $13.8 billion valuation, is expected to ship product through March. The cadence of recent *open-weight* releases (**Mistral** Large 3 in December 2025, Ministral 3 family in late 2025) suggests at least one more flagship **Mistral** release before the end of Q1.

Replit's *Agent* line is expected to receive a meaningful upgrade during March. The pattern of competing autonomous-coding tools — **Cursor** background agents, *Claude Code* subagents, Devin's *Agent* Compute Units — suggests parallel-task execution will become a baseline expected feature rather than a differentiator.

The *Digital Omnibus on AI* is expected to reach political agreement during March. The legal effect of any postponement of high-risk obligations will then depend on a formal vote during April or May. Companies operating in or selling into the European market should treat the 2 August 2026 deadline as binding until any postponement is enacted.

Q1 venture-deployment data will close during March, with the **OpenAI** $122 billion round expected to be confirmed final by quarter-end. Reported AI share of total venture deployment exceeded 80% in independent estimates.

No decisive rulings are expected in the principal generative-AI copyright cases during March. Initial summary-judgement rulings remain on track for Q2 2026.