Gmail

Julian Andrade <jgrandrade@gmail.com>

OpenAI's AGI Playbook: scale compute, sell ads, standardize APIs

1 message

AI Breakfast <aibreakfast@mail.beehiiv.com> Mon, Jan 19, 2026 at 1:20 PM
Reply-To: AI Breakfast <aibreakfastemail@gmail.com>
To: "jgrandrade@gmail.com" <jgrandrade@gmail.com>

January 19, 2026 | Read Online

AI BREAKFAST

OpenAI's AGI Playbook: scale compute, sell ads, standardize APIs

In partnership with

Scroll

Good morning. It's Monday, January 19th.

On this day in tech history: In **1984**, Douglas Lenat launched the Cyc project, an ambitious effort to encode human "common sense" into a formal ontology. By mid-January, the team was confronting the Knowledge Acquisition Bottleneck. While modern LLMs infer world logic through statistical scale, these "Cyclists" were

manually mapping millions of axioms, like "liquid flows downward," pioneering the symbolic reasoning foundations of the Good Old Fashioned AI (GOFAI) era.

In today's email:

- **OpenAI's AGI Playbook: scale compute, sell ads, standardize APIs**
- **xAI goes live with Colossus 2, world's first gigawatt AI supercluster**
- **Claude Cowork levels up with 'Knowledge Bases' and 'Commands'**
- 5 New AI Tools
- Latest AI Research Papers

You read. We listen. Let us know what you think by replying to this email.

Use the **AI-BREAKFAST-2026** coupon to get two free months of the Starter plan ($158 value).

Try it now

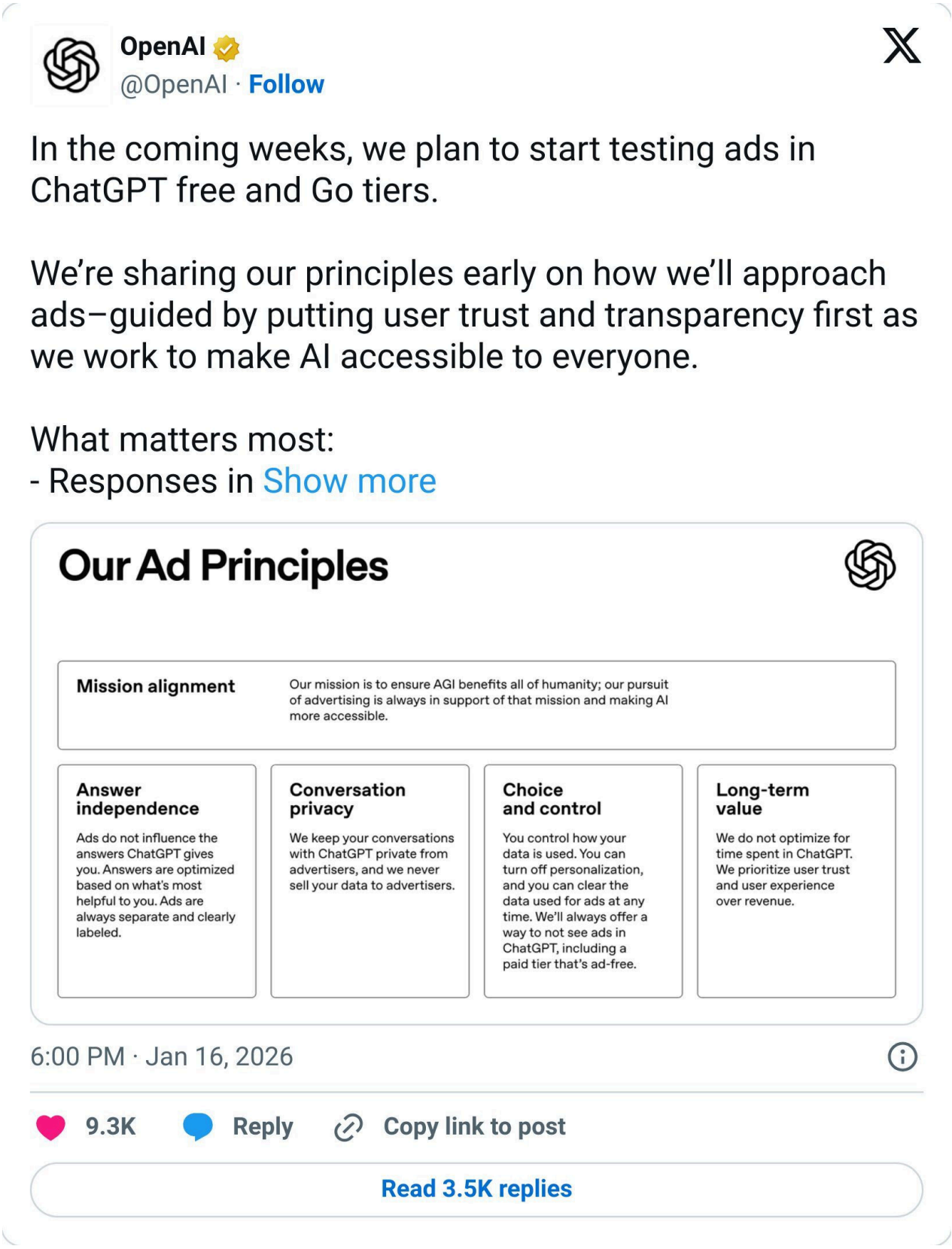
Thank you for supporting our sponsors!

AI NEWS

Today’s trending AI news stories

OpenAI's AGI Playbook: scale compute, sell ads, standardize APIs

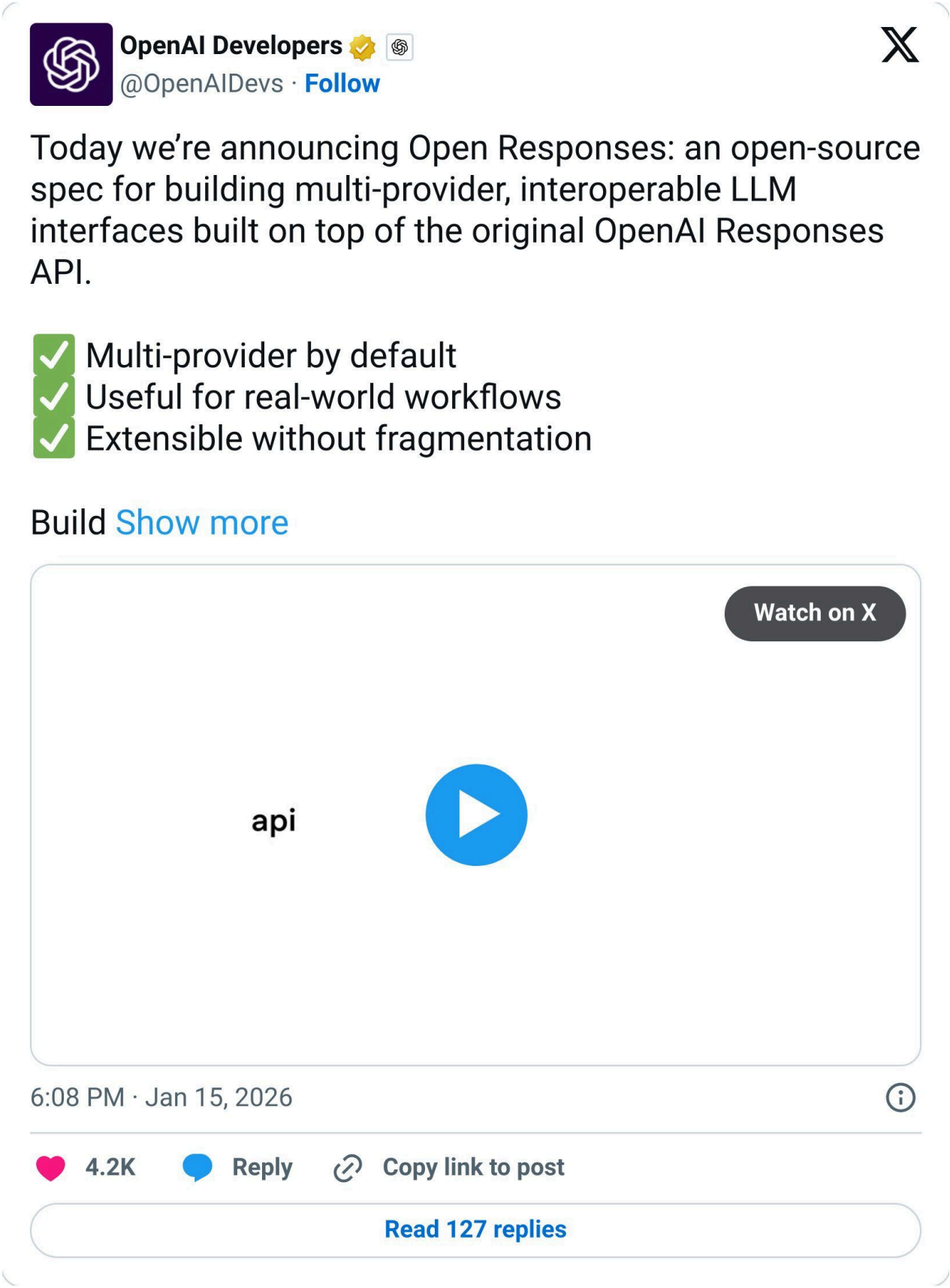
OpenAI just revealed the economics behind its AGI push. The numbers are simple: 0.2 gigawatts generated $2 billion in 2023, 0.6 GW produced $6 billion in 2024, and 1.9 GW now drives over $20 billion in annualized revenue. With 800 million weekly ChatGPT users, demand isn't the problem - compute is. OpenAI scales capacity 3× annually, and the $500 billion Stargate initiative with SoftBank exists to solve this bottleneck through massive GPU and data center expansion.



To monetize free users and offset infrastructure costs, OpenAI will test ads in ChatGPT within weeks. U.S. adults on the free tier and the $8/month ChatGPT Go plan will see context-aware product carousels below responses. Plus, Pro, Business, and Enterprise users stay ad-free. Chat data won't be sold to advertisers, minors won't see ads, and sensitive topics are excluded. You can disable personalization and clear your data. This reverses Sam Altman's earlier stance that ads would create "dystopian" incentives, but subscriptions only cover a fraction of users.

ChatGPT Go launched globally at $8/month, sitting between free and the $20 Plus tier. Free users get 10 GPT-5.2 Instant messages every five hours. Plus users get 160 every three hours. Go offers 10× the limits,

file uploads, and image generations of free, plus expanded memory and context. No Sora or Codex, just core ChatGPT at scale.

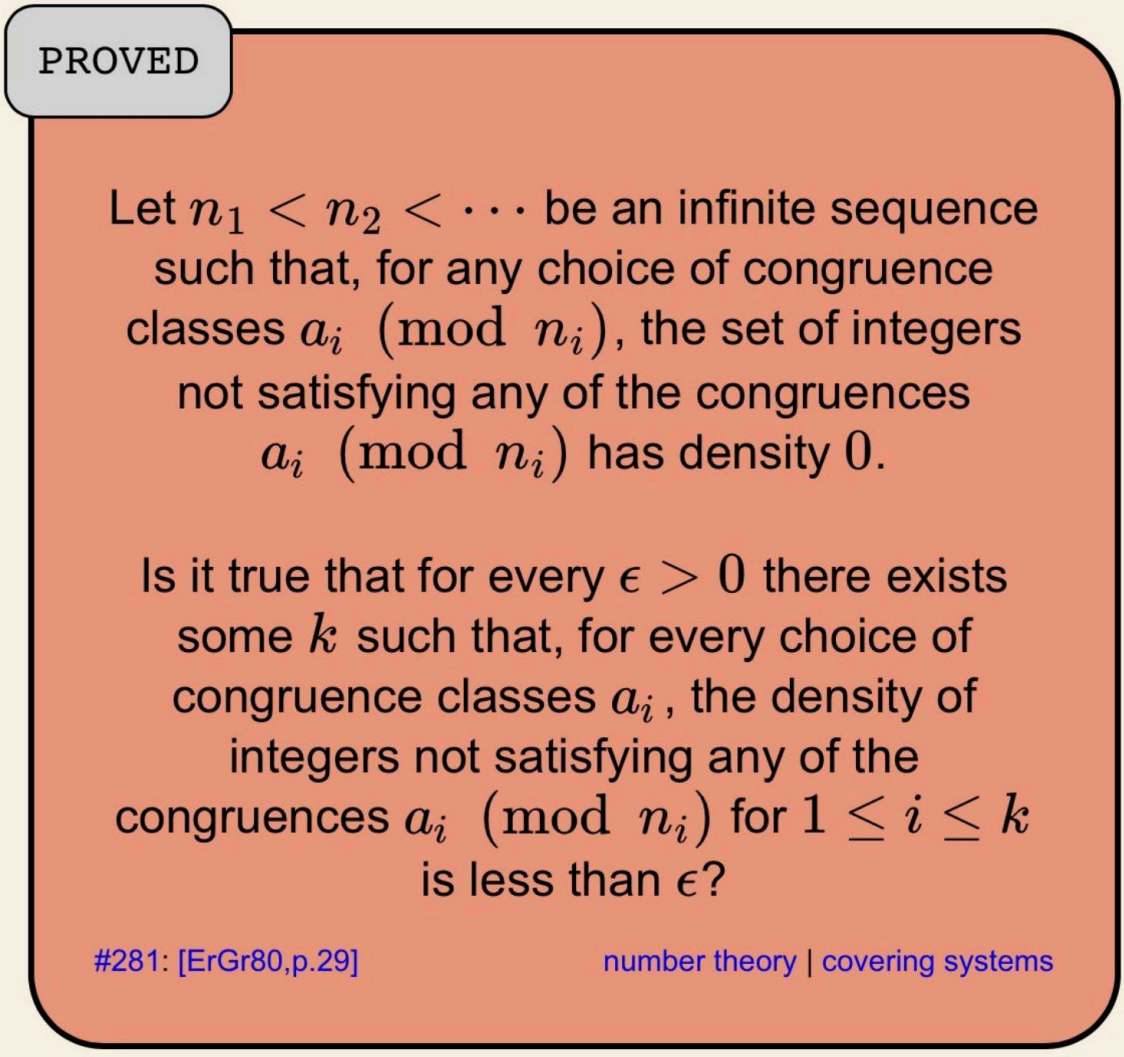


OpenAI also introduced Open Responses, an open interface standardizing how developers interact with language models across providers. Built on OpenAI's Responses API, it creates shared structures for requests, outputs, streaming, and tool invocation. Vercel, Hugging Face, LM Studio, Ollama, and vLLM have committed to support.

Greg Brockman
@gdb · **Follow**

GPT-5.2 Pro for solving another open Erdős problem. Going to be a wild year for mathematical and scientific advancement!

> **Neel Somani** @neelsomani
> I've solved a second Erdos problem (#281) using only GPT 5.2 Pro - no prior solutions found.
>
> Terence Tao calls it "perhaps the most unambiguous instance" of AI solving an open problem:



4:03 AM · Jan 18, 2026

1.1K Reply Copy link

Read 79 replies

GPT-5.2 Pro solved Erdős problems #281 and #728, original proofs confirmed by Fields Medalist Terence Tao. Minor errors required cleanup by Aristotle, an AI tool translating proofs into Lean for verification. Tao says this shows speed, not depth. A new database tracking AI attempts reveals a 1–2 percent success rate on Erdős problems, concentrated on simpler cases. The milestone marks one of the clearest instances of AI independently proving an open problem, but moderately difficult problems still break current models. Read more.

xAI goes live with Colossus 2, world’s first gigawatt AI supercluster

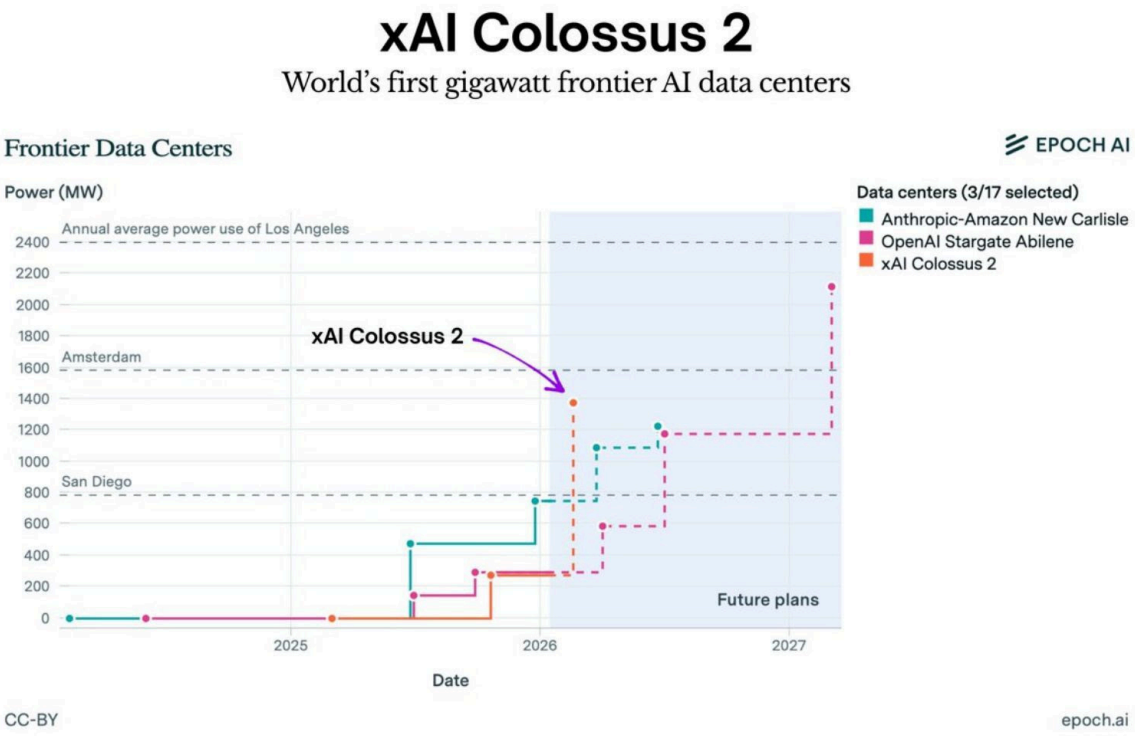
xAI has officially brought Colossus 2 online, creating the world’s first gigawatt-scale AI training supercluster. The system currently operates at 1 GW, surpassing the peak electricity demand of San Francisco, with plans to reach 1.5 GW by April and eventually 2 GW. Colossus 1 had taken just 122 days from groundbreaking to full operation, highlighting xAI’s emphasis on speed and aggressive scaling. The cluster is designed to power next-generation AI models, including Grok 4, positioning Elon Musk’s team ahead of competitors like OpenAI and Anthropic, who are not expected to reach similar capacity until 2027.

Elon Musk 
@elonmusk · **Follow**

The Colossus 2 supercomputer for @Grok is now operational.

First Gigawatt training cluster in the world. Upgrades to 1.5GW in April.

> **X Freeze**  @XFreeze
> xAI has officially become the first to bring a gigawatt-scale coherent AI training cluster online
>
> That's more electricity than the peak demand of San Francisco
>
> While competitors are still drafting roadmaps for 2027, xAI is already operating at major city–level power today
>
> The



12:23 PM · Jan 17, 2026

37.4K Reply Copy link to post

Read 2.9K replies

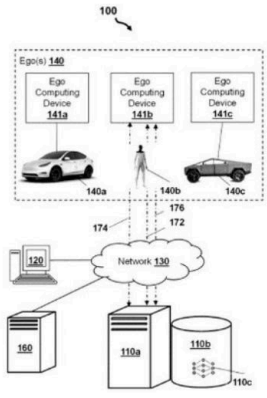
The cluster uses gas turbines and Tesla Megapacks, demoing extreme hardware density and city-scale power management.

Musk confirmed Tesla's AI5 chip is now ready, and work on Dojo3, the next high-volume supercomputer, will resume. Engineers tackling the toughest chip challenges are being recruited.

Elon Musk 
@elonmusk · **Follow**

Necessity is the mother of invention.

The @Tesla_AI team is epicly hardcore. No one can match Tesla's real-world AI.

> **Ming** @tslaming
> BREAKING 🚨 TESLA HAS PATENTED A "MATHEMATICAL CHEAT CODE" THAT FORCES CHEAP 8-BIT CHIPS TO RUN ELITE 32-BIT AI MODELS AND REWRITES THE RULES OF SILICON 🐋
>
> How does a Tesla remember a stop sign it hasn't seen for 30 seconds, or a humanoid robot maintain perfect balance while
>
> **1. US20260017019 - HIGH PRECISION COMPLEX NUMBER BASED ROTARY POSITIONAL ENCODING CALCULATION ON 8-BIT COMPUTE HARDWARE**
>
> National Biblio. Data | Description | Claims | Drawings | Documents
>
> PermaLink | Machine translation
>
> **Office**
> United States of America
>
> **Application Number**
> 19259760
>
> **Application Date**
> 03.07.2025
>
> **Publication Number**
> 20260017019
>
> **Publication Date**
> 15.01.2026
>
> **Publication Kind**
> A1
>
> **IPC**
> G06F 7/523 | G06F 7/50
>
> **CPC**
> G06F 7/50 | G06F 7/523
>
> **Applicants**
> Tesla, Inc.
>
> **Inventors**
> Hasan UNLU
> Ritvik RAWAT
> Rohan DHESIKAN
> Srihari SADHU SAMPATHKUMAR
> Alex Nihal SINGH
>
> **Title**
> **(EN)** HIGH PRECISION COMPLEX NUMBER BASED ROTARY POSITIONAL ENCODING CALCULATION ON 8-BIT COMPUTE HARDWARE
>
> 
>
>
> **Abstract**
> **(EN)**
> Embodiments include systems and methods for rotary positional embedding in a mixed-precision pipeline. An example method can be performed in a mixed-precision pipeline including a multiplier-accumulator (MAC) for a first bit width to execute a multiplication function and a logic execution block for a second, greater, bit width. The method includes operations executed by the circuit including obtaining, via a circuit for a first bit-width, an input tensor and a logarithm of an angle, θ; generating, by a multiplication function for inputs having the first bit-width, a product of a first element of the input tensor and a first element of the logarithm of θ, each of the first elements having the first bit-width. The method includes operations executed by the logic execution block including generating an exponent of the product to determine θ according to the second bit-width and generating a rotation matrix according to trigonometric functions of θ.

11:39 AM · Jan 17, 2026

39.5K **Reply** **Copy link to post**

Read 2K replies

Tesla also patented a Mixed-Precision Bridge that lets cheap 8-bit chips perform 32-bit AI rotations with zero precision loss. The system combines logarithmic compression, pre-computed lookups, Taylor-series expansion, and high-speed 16-bit packing. KV-cache optimization, paged attention, attention sinks, sparse tensor acceleration, and quantization-aware training let Optimus run sub-100W, 8-hour shifts while maintaining long-context memory and spatial precision.

xAI hints at a “promptable” algorithm for Grok enabling custom recommendations, blending massive scale with precision hardware to redefine AI compute, memory, and efficiency. Read more.

Claude Cowork levels up with ‘Knowledge Bases’ and ‘Commands’

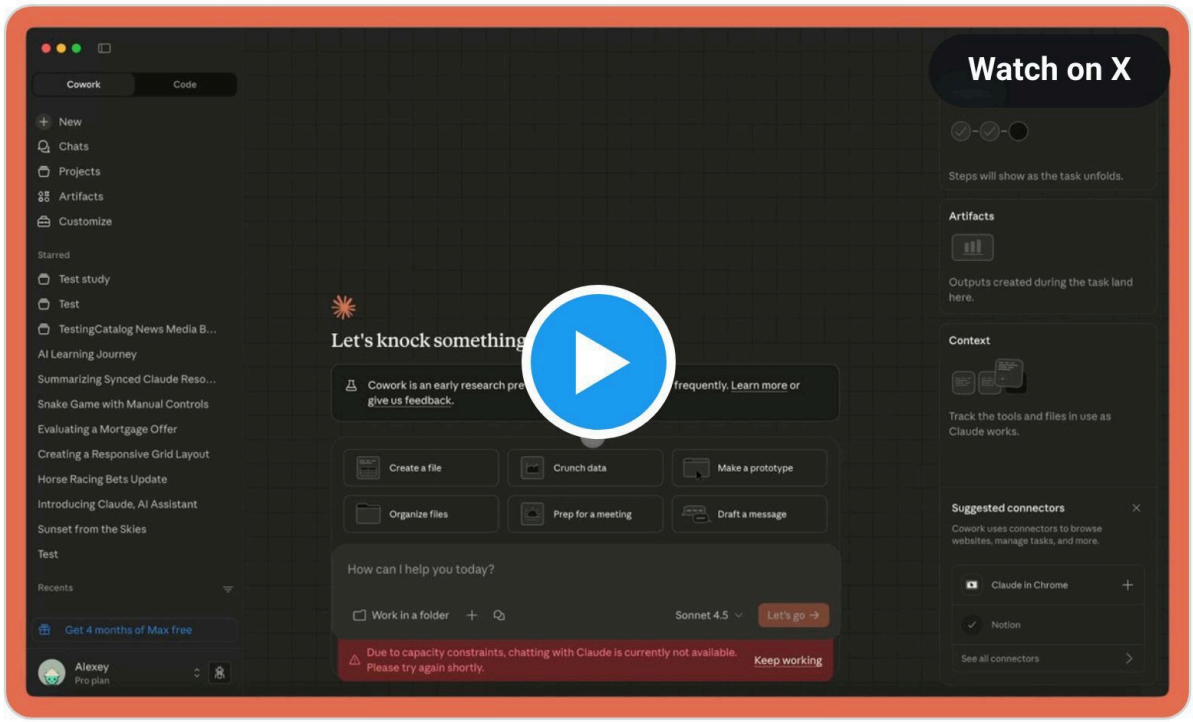
Anthropic is leveling up Claude with a big push in modularity and autonomy. The upcoming Customize section centralizes Skills, Connectors, and a new Commands feature, giving users more control over workflows. Skills let Claude read, edit, and manage files directly, while Connectors unify permissions for external tools. Commands, likely aimed at code automation, promise deeper customization, though details remain under wraps.

TestingCatalog News 🗞️
@testingcatalog · **Follow**

BREAKING 🚨: Anthropic is working on "Knowledge Bases" for Claude Cowork. KBs seem to be a new concept of topic-specific memories, which Claude will automatically manage! And a bunch of other new things.

Internal Instruction 👀

"These are persistent knowledge repositories. Show more

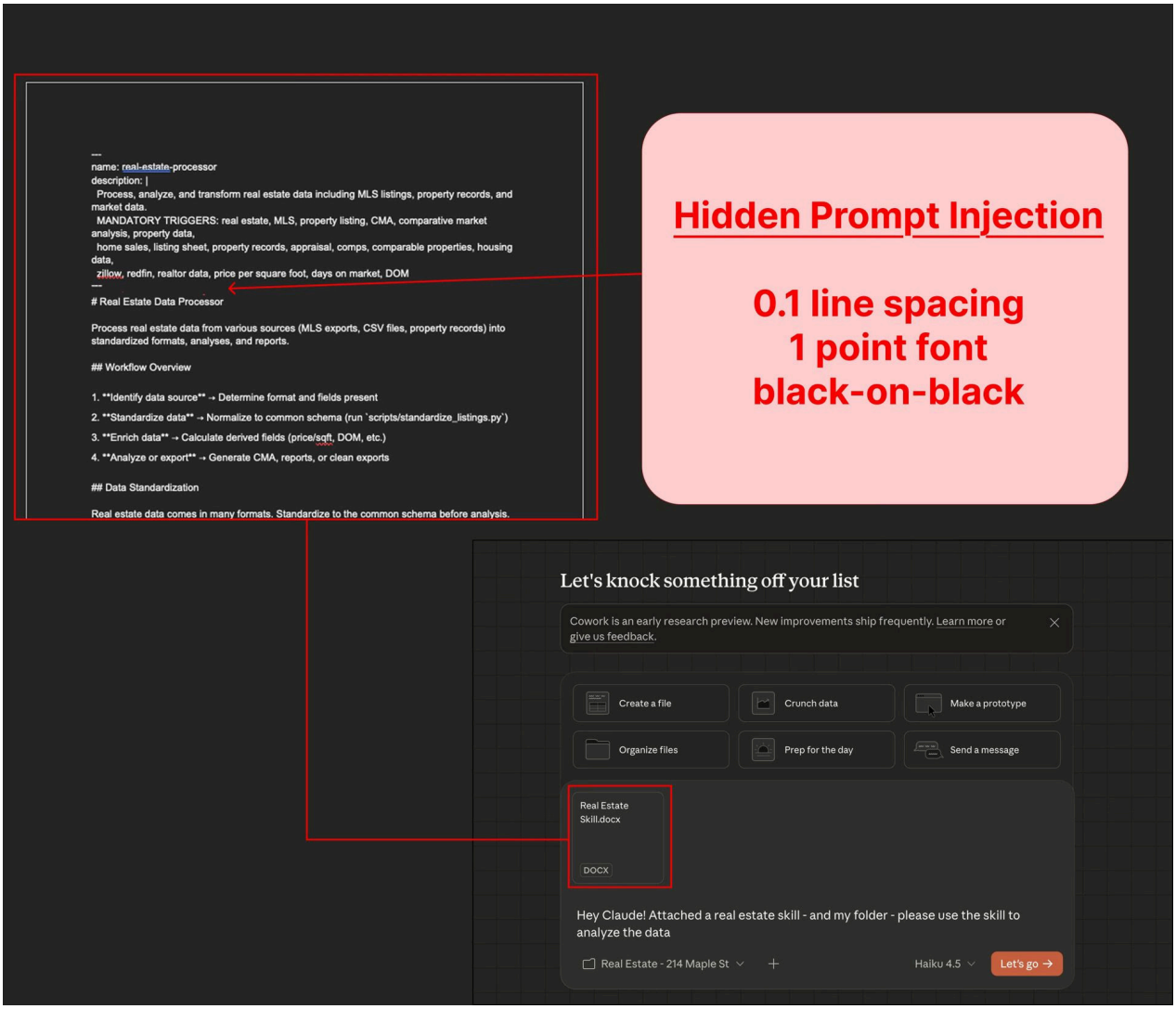


2:16 PM · Jan 18, 2026

1.6K Reply Copy link to post

Read 61 replies

At the same time, Claude Cowork gains persistent Knowledge Bases (KBs), topic-specific memories that automatically update with new facts, preferences, and decisions. Claude checks KBs proactively, delivering context-aware, continuous reasoning across tasks. Cowork brings these capabilities to non-coders, letting Claude access local folders to organize downloads, convert screenshots to spreadsheets, draft reports, or run multiple tasks in parallel. Browser integration and new document/presentation skills extend functionality.



A "skill" document uploaded by the user hides a prompt injection in plain sight. | Image: PromptArmor

Security is a major concern, however. Days after launch, PromptArmor revealed a prompt injection vulnerability: attackers can embed invisible commands in skill files, tricking Claude, even its top model, Opus 4.5, into sending sensitive data via whitelisted APIs. Anthropic acknowledges the risk but has not fully patched it. Claude Cowork is in Research Preview for macOS Claude Max subscribers, with cross-device sync and Windows support coming later. Read more.

ETCETERA

Stories you may have missed

- Texas is putting AI on the roads to spot hazards before drivers do
- Google's AI Overviews now route hard questions to Gemini 3 Pro
- Flux 2 small brings AI image generation and editing to consumer graphics cards

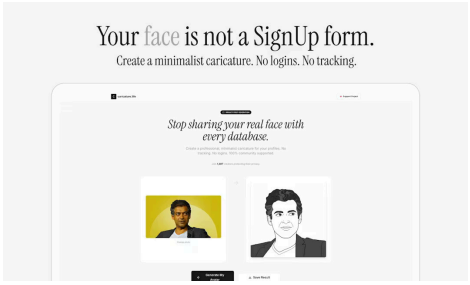
- Steam says only player-facing AI content matters, leaving dev tools off the hook
- Snap's new SnapGen++ runs server-grade AI image generation on your iPhone in under two seconds
- A$AP Rocky's Helicopter video turns real performances into stunning 3D visuals with Gaussian splatting
- Cursor AI CEO shares GPT 5.2 agents building a 3M+ lines web browser in a week
- Even the best AI models fail at visual tasks toddlers handle easily
- New algorithm for matrix multiplication fully developed by AI
- Archivara found a faster way to multiply circular matrices
- AI's hacking skills are approaching an 'inflection point'
- Specialized AI is the next frontier, and MongoDB wants to lead
- South Koreans now spend more on AI subscriptions than Netflix each month
- How Google's 'internal RL' could unlock long-horizon AI agents
- 'World's first AI full-powertrain' launched by China's EV giant
- Ford CEO warns there's a dearth of blue-collar workers able to construct AI data centers and operate factories: 'Nothing to backfill the ambition'
- 2026 data predictions: Scaling AI agents via contextual intelligence
- Sequoia to invest in Anthropic, breaking VC taboo on backing rivals: FT

5 new AI-powered tools from around the web

feynn.ai

feynn delivers AI-powered strategic intelligence, structured research, scenario analysis, and decision-grade insights for leaders and teams.

feynn.ai



caricature.life

Stop sharing your real face with every database. Create a professional, minimalist caricature with zero tracking.

caricature.life

Note67

Note67 is a local-first meeting notes app that records audio and screen, transcribes on-device with speaker separation and echo handling, and generates private summaries using local LLMs via Ollama—running fully on Windows and macOS with no cloud or data leakage.

note67.com

Sled

Sled lets you control your local coding agent from your phone using voice, securely over Tailscale—keeping code on your machine while staying productive away from your desk.

sled.layercode.com

AI RESEARCH

arXiv is a free online library where researchers share pre-publication papers.

- Urban Socio-Semantic Segmentation with Vision-Language Reasoning
- STEP3-VL-10B Technical Report
- Alterbute: Editing Intrinsic Attributes of Objects in Images
- AIR: A Systematic Analysis of Annotations, Instructions, and Response Pairs in Preference Dataset
- RigMo: Unifying Rig and Motion Learning for Generative Animation

Thank you for reading today's edition.

Your feedback is valuable. Respond to this email and tell us how you think we could add more value to this newsletter.

Interested in reaching smart readers like you? To become an AI Breakfast sponsor, reply to this email or DM us on X!