DeepMind's Math Breakthrough, xAI's Grok Sprint, and the Compute Squeeze

AI High Signal Digest

2026-05-25

DeepMind's Math Breakthrough, xAI's Grok Sprint, and the Compute Squeeze

By AI High Signal Digest • May 25, 2026

DeepMind reported a formally verified math breakthrough, xAI shipped Grok 4.3 while preparing a larger V9 model, and multiple signals pointed to compute becoming the core constraint in frontier AI. Also in the brief: long-context training research, new agent tooling, and mixed labor signals as AI adoption broadens.

Top Stories

Why it matters: the strongest signals today were verified reasoning gains, faster frontier model iteration, and growing pressure around compute access.

- **DeepMind reported a formally verified math advance.** AlphaProof Nexus solved 9 open Erdős problems, some unsolved for 56 years, and also proved 44 open OEIS conjectures, resolved a 15-year-old algebraic geometry question, and found a novel optimization parameter [1]. The system combines LLM reasoning with Lean verification, and one analysis said a simple generate-check loop matched the full system on all nine Erdős successes, underscoring how formal verification can filter hallucinations in hard reasoning tasks [1].
- **xAI is compressing its model cycle.** Grok 4.3 is now live on the xAI API, with a 1M-token context window, pricing of $1.25/m input and $2.50/m output, and leaderboard claims in tool use, instruction following, and enterprise domains [2]. Separately, xAI said Grok V9-Medium (1.5T) finished training, with fine-tuning underway, reinforcement learning starting in days, and public release targeted in 2-3 weeks; Elon Musk said it should materially improve harder coding tasks over the current production model [3].

- **Compute pressure is intensifying.** GPU rental prices are up more than 2x since January 2026 [4], while one prominent view this week was that critical-path AGI pretraining now effectively requires the compute scale of OpenAI, Google, Meta, or the Anthropic/xAI/Cursor group [5]. Against that backdrop, Meta cutting 8,000 jobs while spending $100 billion on AI data centers stood out as a stark capital-allocation signal [6].

Research & Innovation

Why it matters: the most useful research updates were about training models more efficiently and measuring their behavior more honestly.

- **Long-context pretraining still has architectural traps.** An AllenAI/CMU paper found 4k-token pretraining metrics have little correlation with actual long-context performance, and recommended avoiding QK norm, Group Query Attention, and Sliding Window Attention while pretraining on longer sequences [7, 8]. Paper: allenai.org/papers/olmpool [9].
- **OPUS moves data selection from static to dynamic.** The ICML Oral paper dynamically selects training data at every pretraining iteration and reported better efficiency and model quality than static selection across language tasks [10].
- **A large behavior study raised another warning on post-training.** Testing models on data from more than 200,000 participants and nearly 26 million human responses, the authors found post-training made models less human-like; related commentary warned that optimizing narrow objectives can shift behavior in unrelated domains [11].

Products & Launches

Why it matters: launches centered on enterprise deployment, agent infrastructure, and faster local inference.

- **Cohere open-sourced Command A+.** The 218B/25B-active MoE targets enterprise agentic workflows, adds multimodal reasoning, supports 48 languages, and can run on as little as two H100s or one Blackwell GPU [12].
- **Cloudflare expanded Think for agent orchestration.** New updates add support for the agentskills.io spec, local/codebase/R2 skill loading, a configurable permission model, and JS/Python/Bash scripts with workspace access; scheduled tasks can run prompts on cron patterns or a DSL [13, 14].
- **Local inference got faster.** llama.cpp with MTP support pushed Qwen3.6-27B dense generation on an A10G from 25 tok/s to 45 tok/s, a

78% jump that was framed as making local models more viable as daily drivers [15].

Industry Moves

Why it matters: the business story was split between workforce disruption, expanding software demand, and clearer production use cases.

- **The labor signal remains mixed.** Meta, Cisco, and Intuit were cited cutting 8,000, 4,000, and 3,000 jobs respectively, with over 100,000 tech jobs gone so far in 2026; one analysis argued companies are now more openly shifting spend from headcount to GPU clusters [6].
- **But AI coding may be expanding software demand rather than shrinking it.** David Sacks said software-engineer postings are rising as GitHub commits grow 14x YoY and AI lowers the cost of writing code, enabling more bespoke software across businesses [16].
- **AI video crossed another adoption threshold.** Kling is now being used in TV and film production, and *House of David* was described as the first Hollywood production to openly discuss AI video generation at industrial scale; the show reportedly reached 44M+ viewers and hit #1 on Prime Video U.S. [17].

Quick Takes

Why it matters: a few smaller updates sharpened the picture on security, local AI, and semiconductor competition.

- TrapDoor hit npm, PyPI, and Crates.io with 34 malicious packages and also used poisoned `CLAUDE.md` and `.cursorrules` files to target developers using AI coding tools [18].
- Gemma 4 has been downloaded more than 120 million times just weeks after release [19].
- Hugging Face said 300,000 AI builders completed hardware profiles, another data point behind the rise of local AI [20].
- Huawei claimed a new path to narrow its semiconductor gap with TSMC without cutting-edge equipment [21].

Sources

1. X post by @kimmonismus
2. X post by @xai
3. X post by @elonmusk
4. X post by @AnjneyMidha
5. X post by @_aidan_clark_
6. X post by @kimmonismus

7. X post by @gabriberton
8. X post by @gabriberton
9. X post by @gabriberton
10. X post by @jiqizhixin
11. X post by @ValerioCapraro
12. X post by @dl_weekly
13. X post by @threepointone
14. X post by @threepointone
15. X post by @ClementDelangue
16. X post by @DavidSacks
17. X post by @kimmonismus
18. X post by @kimmonismus
19. X post by @osanseviero
20. X post by @ClementDelangue
21. X post by @business