SANA-WM Arrives as Cyber Capability Curves Steepen and AI Infrastructure Tightens

AI High Signal Digest

2026-05-18

SANA-WM Arrives as Cyber Capability Curves Steepen and AI Infrastructure Tightens

By AI High Signal Digest • May 18, 2026

NVIDIA's open world model was the headline release, while UK AISI signaled faster autonomous cyber progress and new notes pointed to power and GPU supply as real constraints. Also inside: fresh agent research, Hermes and Codex product updates, and new hardware strategy signals from China, Cerebras, and open-model advocates.

Top Stories

Why it matters: today's biggest signals were a major open world-model release, a sharper cyber capability warning, and mounting infrastructure strain.

- **NVIDIA released SANA-WM.** The 2.6B-parameter open-source world model generates controllable 720p videos up to 60 seconds from one image, a text prompt, and a 6-DoF camera trajectory. It is described as running locally on a single RTX 5090-class GPU, denoising a full 60-second clip in about 34 seconds, with 36× higher throughput than earlier open models [1, 2].
- **The UK AI Safety Institute flagged a faster cyber-capability curve.** It said the length of cyber tasks frontier models can autonomously complete is doubling every 4.7 months, versus 8 months last November, and that Claude Mythos Preview and GPT-5.5 are already above that trend [3].
- **Power and GPU supply look increasingly constraining.** One note said the proposed Stratos data center in Utah could consume up to 9 GW at full buildout, roughly New York City's average electricity demand, while another said H100s now cost more than they did three years ago and remain unavailable on demand because large labs have locked up supply

[4, 5].

Research & Innovation

Why it matters: the most useful research today focused on better ways to reason, search, and train agents under real constraints.

- **On Training in Imagination** separates dynamics error from reward error in model-based RL under imperfect world models and limited budgets. The reported takeaways: reward models scale faster with data than dynamics models, smoother low-Lipschitz models produce more stable rollouts, and many cheap noisy reward labels can outperform fewer accurate ones, though biased rewards are especially risky [6].
- **OpenDeepThink** scales test-time compute through parallel populations of candidate solutions instead of a single longer reasoning trace. In competitive programming, it improved Gemini 3.1 Pro by +405 Codeforces Elo across eight sequential LLM-call rounds [7, 8].
- **Is Grep All You Need?** argues agent harness design matters as much as retrieval. Across LongMemEval tasks, grep-style search beat vector retrieval, especially for coding-style evidence-location problems such as finding exact symbols, diffs, or failing tests [9].

Products & Launches

Why it matters: product updates centered on making agents more useful in day-to-day workflows.

- **Hermes Agent v0.14.0** added xAI SuperGrok and Premium+ access for Grok models, image and video generation, X search, a Codex backend for OpenAI models, a LINE gateway, native video generation, and a Windows native beta [10, 11].
- **Codex appears to be moving into broader desktop workflows.** A recent demo showed agentic Excel on Mac, alongside roadmap hints from a keynote and a draft guide from a Codex team member on daily-use primitives [12, 13, 14].
- **Anthropic released a two-hour training on building Claude agents.** The course covers unsupervised agent structure, terminal access, file-system memory, hallucination-blocking hooks, and operating on large codebases more safely [15].

Industry Moves

Why it matters: hardware access and open-model strategy are becoming strategic levers, not just engineering choices.

- **China is planning a large AI token-factory buildout in Wuxi.** The initial deployment uses four Huawei CloudMatrix 384 systems and

was described as the largest token factory in China; one estimate put it at roughly 1.5K H800s and 3 million V3 tokens per second [16, 17].

- **OpenAI's Cerebras interest was framed as a timeline decision.** In trial testimony, Greg Brockman said he and Ilya Sutskever estimated AGI would take 15 years on standard computing progress, but Cerebras hardware could cut that to 5 years, which he said is why OpenAI explored a merger with Cerebras [18].
- **The open-model geopolitical debate sharpened.** One analysis warned that without a credible Western open frontier player, Chinese open models could become the default across much of the world by 2030; Yann LeCun pointed to Project Tapestry as the response [19, 20].

Quick Takes

Why it matters: several smaller updates still highlighted reliability, security, and adoption shifts.

- Fine-tuning on documents that explicitly say an implausible claim is false can still make models believe the claim; the issue was noted in GPT-4.1 and Kimi K2.5 [21, 22].
- KV cache flushing in Claude Code appears to degrade performance; a related note says KV states carry information that text tokens alone do not, so flushing can reduce accuracy [23, 24].
- OpenAI said ChatGPT Images 2.0 has already generated more than 1 billion images in India [25].
- A recent TanStack supply-chain attack was described as specifically targeting AI developer tooling [26].

Sources

1. X post by @BrianRoemmele
2. X post by @matvelloso
3. X post by @dl_weekly
4. X post by @kimmonismus
5. X post by @Yuchenj_UW
6. X post by @TheTuringPost
7. X post by @wenhaocha1
8. X post by @teortaxesTex
9. X post by @rohanpaul_ai
10. X post by @NousResearch
11. X post by @Teknium
12. X post by @swyx
13. X post by @swyx
14. X post by @jxnlco
15. X post by @Jouhatsu_ai

16. X post by @harukaze5719
17. X post by @teortaxesTex
18. X post by @MTSlive
19. X post by @Dan_Jeffries1
20. X post by @ylecun
21. X post by @OwainEvans_UK
22. X post by @paul_cal
23. X post by @DimitrisPapail
24. X post by @teortaxesTex
25. X post by @sama
26. X post by @thursdai_pod