Open-Weight DeepSeek Nearly Matches OpenAI's Restricted Cyber Model at Breaking Into Enterprise Networks

AI High Signal Digest

2026-11-07

Open-Weight DeepSeek Nearly Matches OpenAI's Restricted Cyber Model at Breaking Into Enterprise Networks

By AI High Signal Digest • November 7, 2026

Vals AI's CyberOffense benchmark finds that a freely downloadable model breaks into simulated corporate networks at close to frontier level. Also this period: Thinking Machines' Inkling 2 Small posts a big ARC-AGI result, AI-assisted math and cryptanalysis results land, and the Nvidia–Megaspeed spy reporting sharpens the export-control fight.

Frontier-level cyber offense is now available as open weights

Vals AI released **CyberOffense**. In the benchmark, each agent works alone from an isolated attack machine against two simulated enterprise networks. It gets a public web address and no human help. Vals says the results show agents can now break into realistic corporate networks unaided, and that this capability is "available to anyone" [1].

- **Scores.** GPT-6 Sol scored best at 80.8% of objectives. It is only available through OpenAI's restricted Daybreak cyber program. **DeepSeek V4.1 Flash 1017**, an open-weight model, scored 72.8% and matched Daybreak on the Nova network. On Lighthouse, Daybreak completed 27 of 31 objectives in under three hours, and DeepSeek completed 21 [2].
- **Cost and token budget.** A DeepSeek run cost under $3, against about $22 for Daybreak, and the open model ran with no safeguards. DeepSeek completed nothing in its first 25M tokens and made most of its progress after 100M. Vals concludes that short tests understate what open models can do. It also expects attacks like these on real networks to become

more common in the coming months, and urges defenders to use the same models to find and fix weaknesses first [3].

Vals' point is that gated access matters less once open weights are this close behind. OpenAI's restricted Daybreak program, covered in recent briefs, is the clearest example. Google is following the same gated approach: per a newsletter summary, **Gemini 4 Carbon** tops the Artificial Analysis Intelligence Index at 59, largely on cybersecurity gains, but only vetted cyber defenders can use it [4].

DeepMind cryptanalysis. Google DeepMind reports that AI agents, working under human supervision, found two weaknesses in legacy RSA-1024: - **Signatures.** Padding adds no protection once an attacker controls roughly half or more of the signed block. For cases where the attacker controls less, a new algorithm is estimated to improve on the previous best method by about 1,600× [5]. - **Batch factoring.** A new algorithm cuts the average cost of factoring each additional 1024-bit key by about 30× [5].

DeepMind stresses that modern 2048-bit and larger keys remain secure. Its argument is that AI can now probe hardness assumptions much faster than the years or decades such advances used to take [5].

Inkling 2 Small: strong ARC results and signs of a cheap training run

Thinking Machines launched Inkling 2 Small last period. This period brought third-party results: - **ARC-AGI.** ARC Prize verified 81.7% on ARC-AGI-2 at $0.27 per task and 96.5% on ARC-AGI-1 at $0.08. It says the model beats every open-weight model it has evaluated on both [6]. - **Code Arena.** Arena lists it at 1648 points, #8 among labs in Code Arena: WebDev. It would be the #1 US open model once weights ship [7].

The compute behind it is drawing attention. One analyst estimates pretraining used a bit under 1,000 GB300 GPUs and about 130K GPU-hours, roughly $650K at rental rates. That is his estimate, not a disclosed figure [8].

Funding. A VC post describes a "$5B @ $40B" deal. It says Thinking Machines has committed to deploy more than 1 GW of Vera Rubin compute, which the post puts at $50B+, and has over $100M in ARR [9].

Behavior under evaluation. Vals found profanity in some of the model's benchmark runs. It was not a reliable sign of failure: 96% of failed runs never swore, and two looping runs produced 241 of the 316 uses [10].

AI-assisted science: a math proof and two new lab benchmarks

- **Aaronson–Ambainis conjecture.** Physical Superintelligence says its research system, using a non-public Anthropic model, proved this con-

jecture in quantum complexity, open since 2008. The system reportedly found a complete proof on its third day, and the final version was checked in Lean [11].

- **VIALS-AA.** Artificial Analysis launched this leaderboard of 161 visual-reasoning tasks drawn from life sciences, such as gel blots, flow cytometry and plasmid maps. A task counts only if the model answers correctly in all three runs.
  - Gemini 4 Carbon leads at 64%, followed by GPT-6 Astra at 63% and Claude Opus 5.5 at 61%. Kimi K3.1 Pro is the best open model at 45%.
  - Cost per task varies widely. Opus averages $1.64, about 29× Sol's ~$0.056, largely because it takes about 27 turns per task against about 8 for Astra and Sol [12].
- **BioBasicsBench.** OpenAI and Ligo released 100 tasks taken from real experiments. They test whether agents can interpret raw lab data, catch errors, and recognize when evidence does not support a conclusion. OpenAI frames the benchmark partly as tracking biosafety and biosecurity [13].

China, compute and the Nvidia–Megaspeed reporting

"Distillation lag." An X article translates minutes of a private October discussion among Chinese AI insiders. They argue the commonly cited "3–4 months behind" reflected the time Chinese labs needed to distill US models, not underlying capability. Their claims: - Distillation costs have risen 5–10×, and up to 20× in some cases, as US labs ban suspected distillers and hide reasoning traces [14]. - Chinese labs hold the equivalent of 10K to a few tens of thousands of Blackwells, against a projected 2–3 million for a single US lab next year. On that basis the room expects the gap to widen to about a year, or 18–24 months if US labs reach recursive self-improvement [14]. - Observer teortaxesTex calls the distillation thesis "largely vindicated" but doubts the claim that Beijing would back DeepSeek [15].

Huawei capacity. Epoch estimates Huawei's Horinger data center at about 80K H100-equivalents, or 7% of Colossus 2's ~1.1 million [16].

Megaspeed. Commentator Chris McGuire posted a recap of FT, NYT and Bloomberg reporting: - US intelligence assessed that Megaspeed executive Alice Huang was a Chinese spy, and that her goal was to get Jensen Huang to lobby for looser export controls. - Megaspeed is likely a Chinese cutout and remains a preferred Nvidia cloud partner receiving Blackwell chips. - BIS has not tightened controls on Megaspeed or Southeast Asia [17].

These are the recapper's framing of others' reporting. They add pressure to the export-control debate at the same moment Chinese insiders describe compute as the binding constraint.

Baidu. Baidu released open weights for **ERNIE-5.5-Flash-Base**, a 280B-

parameter MoE (14B active) trained on 33.5T tokens. In Baidu's preliminary evaluations it beats DeepSeek-V4-Flash-Base, scoring 78.4 vs 69.2 on MMLU-Pro. Its license excludes the EU, UK and South Korea [18].

Products and infrastructure

- **Meta.** Muse and Meta AI now run specialized models trained with Muse Wonder. They add "Active Memory" and "Active Focus," and Meta claims a 2.4× reduction in forgetting goal-critical information and 67% better recovery when constraints change [19].
- **Anthropic.** An Engineering Blog post describes "dreaming," a process by which Claude organizes and corrects what it has learned so it can find it more easily later [20].
- **GPT Image 2.6.** The models take #1 and #2 on Artificial Analysis' text-to-image and image-editing leaderboards. Flare is almost twice as fast as 2.5, and pricing is unchanged at about $0.21 per max-quality image [21].
- **Vera Rubin inference.** On agentic workloads with DeepSeek V4.1 Flash, NVIDIA's Vera Rubin configurations deliver 5.0× GB300's throughput per GPU at 260 tok/s per agent, but only 1.7× at 80 tok/s per agent [22].
- **GPU market.** SemiAnalysis estimates at most 2–3 GW of the ~50 GW of capacity due online in 2027 is at risk. One-year H100 contract prices are up about 21% since March [23].
- **Anthropic IPO.** Per the WSJ as relayed on X, Anthropic's annualized revenue run rate topped $80B in October. Investors are questioning a $1.5T IPO valuation [24].

Watch

METR follow-ups. Ajeya Cotra says the METR authors' estimate that AI nearly doubles capability progress amounts to researchers experiencing time about 10× faster. A naive extrapolation implies ASI within a year, if bottlenecks don't bind [25, 26]. Ryan Greenblatt says he trusts the ~2× estimate but finds it reasonable to want the redacted evidence [27].

Releases still pending. Qwen 4 was expected this week based on remarks by Alibaba Cloud's CTO [28]. No new frontier model from OpenAI or Anthropic shipped by period's end [29].

Sources

1. X post by @ValsAI
2. X post by @ValsAI
3. X post by @ValsAI
4. X post by @dl_weekly
5. X article by @pushmeet

6. X post by @arcprize
7. X post by @arena
8. X post by @teortaxesTex
9. X post by @armenberjikly
10. X post by @ValsAI
11. X post by @kimmonismus
12. X post by @ArtificialAnlys
13. X post by @OpenAI
14. X article by @odysseus0z
15. X post by @teortaxesTex
16. X post by @EpochAIResearch
17. X post by @ChrisRMcGuire
18. X post by @kimmonismus
19. X post by @AIatMeta
20. X post by @AnthropicAI
21. X post by @ArtificialAnlys
22. X post by @ArtificialAnlys
23. X post by @SemiAnalysis_
24. X post by @kimmonismus
25. X post by @ajeya_cotra
26. X post by @ajeya_cotra
27. X post by @RyanGreenblatt
28. X post by @kimmonismus
29. X post by @kimmonismus