Anthropic's cyber push, OpenAI's agent stack, and a widening open-model race

AI News Digest

2026-04-08

Anthropic's cyber push, OpenAI's agent stack, and a widening open-model race

By AI News Digest • April 8, 2026

Anthropic led the day with a controlled cyber-defense rollout for Claude Mythos Preview, while new safety research underscored how fragile agent systems remain once persistent state is compromised. OpenAI offered its clearest picture yet of AI-native software development and a unified app strategy, as Microsoft and Nvidia deepened the open-model competition.

Security moved from warning to deployment

Anthropic launches Project Glasswing around Claude Mythos Preview

Anthropic said its newest frontier model, Claude Mythos Preview, can find software vulnerabilities better than all but the most skilled humans, and has already uncovered thousands of high-severity issues, including some in every major operating system and web browser [1, 2]. Through Project Glasswing, the company is giving controlled access to defenders instead of releasing the model generally, alongside up to $100M in usage credits and partnerships with organizations including AWS, Apple, Google, Microsoft, and the Linux Foundation [3, 4, 5, 6].

> "Rather than release Mythos Preview to general availability, we're giving defenders early controlled access in order to find and patch vulnerabilities before Mythos-class models proliferate across the ecosystem." [3]

Why it matters: Anthropic is explicitly treating frontier-model cyber capability as a current operational risk, and Dario Amodei described Glasswing as a possible blueprint for handling harder model risks still ahead [7, 8, 9].

New agent-safety work argues the weak point is persistent state

A safety evaluation of OpenClaw-style personal agents with access to Gmail, Stripe, and the local filesystem found baseline attack success rates of 10% to 36.7%; poisoning persistent capability, identity, or knowledge state raised success to roughly 64% to 74%, and the strongest defense still left capability attacks at about 63.8% [10]. The paper argues these failures are structural rather than model-specific and proposes a stricter `proposal -> authorization -> execution` pattern, where actions are only reachable after deterministic policy checks [10].

Why it matters: As models gain more tool access, the center of gravity is moving from prompt-level safety toward authorization, policy, and system design around the agent [10].

OpenAI made its agent stack more legible

Frontier's internal coding experiment makes the harness the story

OpenAI's Frontier team said a five-month experiment produced an internal beta with more than 1 million lines of code and thousands of pull requests using zero human-written code, with no human review before merge [11]. The setup relied on what Ryan Lopopolo describes as harness engineering: sub-minute build loops, observability, specs, skills, and the Symphony orchestration layer for supervising large numbers of coding agents across tickets and repositories; he also cautioned that the work happened in a greenfield repository [11, 12].

Why it matters: The emphasis here is not just on a stronger coding model; it is on the surrounding build system, context, and control layer that make autonomous agent work practical [11].

Brockman says OpenAI is consolidating around a unified app

Greg Brockman said OpenAI is moving focus away from video generation as a separate branch and toward the GPT/reasoning stack, with top priorities now a personal assistant and an AI that can solve hard problems under tight compute constraints [13]. He described a unified app that brings together ChatGPT, Codex, browsing, and computer use, to be rolled out incrementally over the next few months; separately, Sam Altman said Codex has reached 3 million weekly users and that usage limits will reset at each additional million up to 10 million [13, 14].

Why it matters: OpenAI is now describing the model, memory, harness, and action layer as one product surface rather than separate tools [13].

The open-model contest widened beyond chat

Microsoft pushed the retrieval layer forward with Harrier

Microsoft's Bing team open-sourced Harrier, an embedding model that it says ranks #1 on the multilingual MTEB-v2 benchmark, ahead of models based on Gemini, Gemma, Llama, and Qwen [15, 16]. Microsoft says Harrier supports more than 100 languages and 32K inputs, and is built for Bing semantic search and the web-grounding service that powers nearly every major AI chatbot; the company also argues better embeddings improve answer accuracy and make agents more stable across multi-step tasks [16, 17, 18].

Why it matters: Competition is moving deeper into the retrieval and grounding layer that agent products depend on, not just the assistant on top [17, 18].

Nvidia paired a fully open 120B model with a detailed training recipe

Nvidia released Nemotron-3 120B, a fully open model trained on 25 trillion tokens that, according to the notes, roughly matches top closed frontier models from about 18 months ago [19]. The release comes with a 51-page paper detailing the training process and dataset, plus inference techniques including NVFP4 quantization, multi-token prediction, member layers, and stochastic rounding; the NVFP4 version is described as 3.5x faster than Nvidia's BF16 variant and up to 7x faster than comparable open models with similar accuracy [19].

Why it matters: This is a notable signal in the open-model race: major vendors are releasing not just weights, but more of the recipe for how to train and serve them efficiently [19].

Sources

1. X post by @AnthropicAI
2. X post by @AnthropicAI
3. X post by @DarioAmodei
4. X post by @AnthropicAI
5. X post by @AnthropicAI
6. X post by @AnthropicAI
7. X post by @DarioAmodei
8. X post by @DarioAmodei
9. X post by @DarioAmodei
10. r/MachineLearning post by u/docybo
11. Extreme Harness Engineering for Token Billionaires: 1M LOC, 1B toks/day, 0% human code, 0% human review — Ryan Lopopolo, OpenAI Frontier & Symphony
12. Extreme Harness Engineering for the 1B token/day Dark Factory — Ryan Lopopolo, OpenAI Frontier

13. OpenAI President Greg Brockman: Doubling Down on Text Models, The Superapp Plan, Codex's Potential
14. X post by @sama
15. X post by @mustafasuleyman
16. X post by @JordiRib1
17. X post by @mustafasuleyman
18. X post by @mustafasuleyman
19. NVIDIA's New AI Just Changed Everything