Anthropic Holds Back Mythos as OpenAI Pushes Deeper Into Science

AI News Digest

2026-06-18

Anthropic Holds Back Mythos as OpenAI Pushes Deeper Into Science

By AI News Digest • June 18, 2026

Anthropic's CEO explained why the company is withholding Mythos and where it draws red lines on cyber and military use. OpenAI paired a new life-science benchmark with a lab-backed chemistry result, while Noam Shazeer's move to OpenAI and claims around Z.ai's Huawei-trained GLM-5.2 underscored intensifying competition.

The main signal

Today's developments were less about new chat surfaces and more about **where frontier AI is allowed to go**: into cyber operations, into real scientific workflows, and into the talent and hardware stacks that will shape the next competitive cycle.

Safety and strategy at the frontier

Anthropic says Mythos stays limited until cyber safeguards improve

Anthropic CEO Dario Amodei said the company withheld Mythos after seeing a large jump in its ability to find vulnerabilities and turn them into exploits autonomously across the cyber kill chain [1]. He said Anthropic is widening access gradually, starting with defenders, because current cyber safeguards can still be jailbroken and are not yet strong enough for a broad release [1].

> "this is a super weapon … Please don't release this." [1]

Amodei also said Anthropic will support some defense use cases while maintaining red lines against mass surveillance and fully autonomous weapons, with humans retaining the final targeting decision [1].

Why it matters: Anthropic is explicitly tying release policy to both the current limits of jailbreak defenses and a narrower definition of acceptable defense use [1].

AI moves deeper into lab work

OpenAI pairs a life-science benchmark with a chemistry result

OpenAI introduced LifeSciBench, a benchmark built with 173 biotechnology and pharmaceutical scientists that includes 750 expert-authored tasks across seven biological research workflows [2]. The benchmark is meant to test whether models can reason from evidence, work with scientific artifacts, handle uncertainty, and make decisions under real-world constraints; OpenAI said GPT-Rosalind scores above GPT-5.5 across all seven workflows [3].

Separately, OpenAI said GPT-5.4 helped drive a medicinal chemistry project from literature review to a validated result with Molecule.one's Maria AI and a specialized lab [4, 5]. In testing, yields improved for 88% of boronic acids and 83% of sulfonamides, and 11 of 14 hand-validated reactions showed higher yields, including 8 with more than twofold improvement; the full process took about 2.5 months [6, 7].

Why it matters: Taken together, the two announcements connect evaluation and execution: OpenAI is not just publishing a science benchmark, but also pointing to a human-validated chemistry campaign as an early example of models supporting more of the research loop [7, 3].

NVIDIA fills in the operational details behind ENPIRE

NVIDIA's GEAR lab shared new details on how ENPIRE runs unattended robot experiments safely: hard kinematic limits trigger task failure and auto-reset, torque-limited compliant grippers turn bad contact into a safe stall, and reward functions are frozen before AutoResearch begins so agents cannot rewrite their own success criteria [8]. The system also tracks Mean Robot Utilization, Mean Token Utilization, GPU utilization, Tokens-to-Success, and Time-to-Success [8].

Why it matters: This is a practical look at what one lab thinks is required before autonomous experimentation can run overnight on physical hardware [8].

Competition keeps tightening

Noam Shazeer is joining OpenAI

Noam Shazeer said he is joining OpenAI after leaving Google [9]. Sam Altman replied that Shazeer is one of the people he has most wanted to work with since OpenAI's founding, while Nathan Lambert called it a major talent move and joked that OpenAI had fixed its supposed "scaling pretraining problem" [10, 11, 12].

Why it matters: Even without technical details, the public reaction framed this as a strategically important talent gain for OpenAI's model-development effort [11, 12].

GLM-5.2 sharpens the debate over a Chinese AI stack

Artificial Analysis's Intelligence Index published its conclusion on Z.ai's GLM-5.2 release [13]. Emad Mostaque said the model was trained on Huawei Ascend chips with no NVIDIA hardware and described it as running on a fully Chinese stack that is roughly three months behind leading models and 90% cheaper; he also estimated total cost at $25 million, mostly post-training [14].

Why it matters: The notable signal is not just model quality, but the claim that competitive systems can be built on a non-NVIDIA stack, which would matter for both AI economics and geopolitics if it holds up [14].

One useful read for operators

Andrew Ng says the bottleneck is shifting from models to workflow design

Andrew Ng said coding agents are moving unusually fast, with teams now mixing Claude Code, OpenAI Codex, and Gemini CLI, and with more coding happening on phones than he would have expected a year ago [15]. But he argued that enterprise ROI depends less on automating one step and more on redesigning whole workflows—such as compressing loan approval from a week to 10 minutes—and that unstructured data architecture is becoming a major blocker for agent deployment [15].

Why it matters: For teams trying to operationalize agents, Ng's message was simple: model progress is no longer the only constraint; workflow redesign and data readiness are becoming the harder part [15].

Sources

1. Inside the Mind of Anthropic CEO Dario Amodei | The Circuit | Extended Interview
2. X post by @OpenAI
3. X post by @OpenAI
4. X post by @OpenAI
5. X post by @OpenAI
6. X post by @OpenAI
7. X post by @OpenAI
8. X post by @DrJimFan
9. X post by @NoamShazeer
10. X post by @sama

11. X post by @natolambert
12. X post by @natolambert
13. X post by @ZixuanLi_
14. X post by @EMostaque
15. The Future of AI Agents with Andrew Ng | Interrupt 26