Anthropic's Mythos Forces a Safety Pivot as GLM-5.1 Raises the Open-Model Bar

AI High Signal Digest

2026-04-08

Anthropic's Mythos Forces a Safety Pivot as GLM-5.1 Raises the Open-Model Bar

By AI High Signal Digest • April 8, 2026

Anthropic unveiled Project Glasswing around Claude Mythos Preview and withheld the model from general release, signaling a new release pattern for cyber-capable frontier systems. Meanwhile GLM-5.1 pushed the open-model frontier forward, and a new wave of agent, retrieval, and workflow products expanded practical AI adoption.

Top Stories

Why it matters: The biggest shift this cycle is not just better model performance. It is a sharper split between tightly controlled frontier systems and fast-improving open and productized AI tooling.

Anthropic turns Claude Mythos into a restricted cyber-defense program

Anthropic launched Project Glasswing, an initiative to secure critical software powered by Claude Mythos Preview, which it says can find software vulnerabilities better than all but the most skilled humans [1]. Anthropic says Mythos has already found thousands of high-severity vulnerabilities, including some in every major operating system and web browser [2]. The launch group includes AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks [3]. Anthropic is committing up to $100M in Mythos Preview usage credits for partners and more than 40 additional organizations that maintain critical software, including open-source projects [4].

Anthropic also says it does **not** plan to make Mythos Preview generally available yet. Its stated goal is to deploy Mythos-class systems safely at scale only after it

has safeguards that can reliably block the most dangerous outputs, with testing set to begin on an upcoming Claude Opus model [5]. In Anthropic employees' descriptions, Mythos is their most reliable model to date, but also one that creates more alignment risk because its failures can have larger consequences [6, 7, 8].

> "We find it alarming that the world looks on track to proceed rapidly to developing superhuman systems without stronger mechanisms in place." [9]

Impact: Anthropic is signaling a new deployment norm for cyber-capable frontier models: narrow access, defensive partnerships, and safety gating before broad release.

GLM-5.1 raises the bar for open-weight coding models

Z.ai introduced GLM-5.1 as a new open model ranked **#1 in open source** and **#3 globally** across SWE-Bench Pro, Terminal-Bench, and NL2Repo [10]. The official launch emphasizes long-horizon behavior: GLM-5.1 can run autonomously for **8 hours** and refine strategies through **thousands of iterations** [10]. In third-party benchmark comparisons included in the notes, GLM-5.1 was reported at **58.4** on SWE-Bench Pro, ahead of Claude Opus 4.6, GPT-5.4, and Gemini 3.1 Pro [11]. The model is already available on Hugging Face, with docs live and chat rollout following [10].

Impact: Open-weight models are moving closer to the center of practical coding workflows, not just serving as lower-cost alternatives.

OpenAI's internal software factory is becoming easier to see

OpenAI's developer account said a small team steering Codex opened and merged **1,500 pull requests** to ship a product used by hundreds of internal users with **zero manual coding** [12]. Separately, a Latent Space episode featuring OpenAI engineer `_lopopolo` described a larger internal setup—Frontier and Symphony—operating at **1M lines of code**, **1B tokens/day**, with **0% human code** and **0% human review** before merge [13]. On the adoption side, Codex has reached **3 million weekly users**, and OpenAI says it will reset usage limits at every additional 1 million users until 10 million [14, 15].

Impact: AI-native software engineering is moving from isolated demos to both internal production processes and mass-market developer usage.

Microsoft is treating retrieval as core AI infrastructure

Microsoft open-sourced **Harrier**, an embedding model ranked **#1** on the multilingual MTEB-v2 leaderboard [16]. Microsoft says Harrier supports **100+ languages**, handles inputs up to **32K**, and powers Bing's next-generation semantic search and web-grounding services for AI agents [16]. Mustafa Suleyman

said better embeddings improve retrieval quality, multilingual performance, and the stability of multi-step agent behavior [17, 18].

Impact: The agent stack is not just about better base models. Grounding, retrieval, and embeddings are becoming competitive layers in their own right.

Research & Innovation

Why it matters: This cycle's technical progress was not only about scale. It also showed advances in memory, inference efficiency, and autonomous discovery.

Mythos case studies point to a large cyber jump

Benchmark summaries shared after the Mythos announcement reported **93.8–93.9%** on SWE-Bench Verified, **77.8%** on SWE-Bench Pro, **82** on Terminal-Bench 2.0, and **181** Firefox exploit-writing successes versus **2** for Claude Opus 4.6 [19, 9]. Summaries of Anthropic's technical report also highlighted a **27-year-old OpenBSD vulnerability** and a **16-year-old FFmpeg bug** that had survived **5 million** automated tests [20]. One explanation of the verification process said results were checked through proof-of-concept code, cross-verification by a second Mythos agent, and final review by human security experts [20].

MemPalace shows how local-first memory systems are maturing

MemPalace, an open-source memory system built with Claude, reported the **first perfect** LongMemEval score at **500/500**, plus **92.9%** on ConvoMem and **100%** on LoCoMo [21]. Its architecture stores conversations locally in a structured "palace," compresses broad personal context into about **120 tokens**, and includes contradiction detection [21]. It runs locally, without cloud dependence, under an MIT license [21].

Flow Map Language Models target much faster text generation

A v2 update to Flow Map Language Models argues for a continuous-flow approach to language modeling that can be distilled into **one-step text generation** [22]. The authors report beating discrete diffusion baselines at roughly **8x speed** [22], plus easier inference-time control over topic, sentiment, grammaticality, and safety [22]. Resources were published alongside the update via a blog and paper [22].

AI agents autonomously designed a new physical structure

In one ScienceClaw × Infinite case study, AI agents built a shared representation across **39 resonators** spanning biology, metamaterials, musical instruments, and Bach chorales, identified an unexplored design gap, and generated a **Hierarchical Ribbed Membrane Lattice** [23]. The best design resonated at

2.116 kHz and showed **nine elastic modes** in the **2–8 kHz** band [23]. The team says the mapping, gap identification, design generation, and validation were carried out without human involvement [23].

Products & Launches

Why it matters: Product launches are moving beyond chat interfaces toward domain workflows, agent runtime infrastructure, and better developer ergonomics.

- **Microsoft Harrier:** Microsoft's search team open-sourced Harrier, a multilingual embedding model built for semantic search and RAG-style grounding in agent workflows [16].
- **Cognition SWE-1.6:** Cognition released SWE-1.6 in Windsurf, saying it matches its Preview model on SWE-Bench Pro while improving behavior through more parallel tool use and less looping. It is free for three months at **200 tok/s**, with a **950 tok/s** paid tier via Cerebras [24, 25, 26, 27].
- **OpenAI Prism Paper Review:** Prism added a workflow for reviewing technical and scientific papers, checking math, derivations, notation, units, section consistency, and whether claims are supported by results, then writing an editable LaTeX review directly into the project [28, 29, 30].
- **AWS S3 Files:** AWS introduced S3 Files, which exposes S3 buckets as a high-performance file system. For agents, that means direct mounted read/write access instead of translating between object-store and file abstractions [31, 32].
- **LangSmith Fleet + Arcade:** LangSmith Fleet now connects to **7,500+ / 8,000+** agent-optimized tools through Arcade, giving agents secure access to systems like Salesforce, GitHub, Zendesk, and Asana [33, 34].

Industry Moves

Why it matters: Capital, hiring, and partnerships are clustering around a few themes: cyber defense, industrial AI, and agent ecosystems.

- **Amazon's Project Prometheus is scaling up:** Reporting summarized in the notes says Jeff Bezos is rapidly expanding Project Prometheus, hiring former OpenAI/xAI leader Kyle Kosic, targeting physical-world AI for aviation and engineering, and planning to raise tens of billions [35, 36].
- **Granola raised a large Series C:** Granola raised **$125M** at a **$1.5B** valuation after **250%** quarterly revenue growth, with plans to push its AI meeting-notes product toward agentic task automation [37].
- **Frontier labs are coordinating on model-copying risk:** OpenAI, Anthropic, and Google are reported to be sharing intelligence through the Frontier Model Forum to detect misuse and prevent Chinese competitors from distilling outputs from advanced models [38, 39].
- **MiniMax is deepening its agent distribution:** MiniMax says it is partnering with NousResearch across product and models, and both firms

say MiniMax M2.7 is already one of the most-used models in Hermes Agent [40, 41].

Policy & Regulation

Why it matters: Government involvement is becoming more operational. Labs are discussing specific offensive and defensive capabilities with states, while public-sector AI programs are moving from concept to deployment.

- **Anthropic is formalizing access controls around Mythos:** Anthropic says Mythos Preview will not be generally available until safeguards can reliably block dangerous outputs, and that it will test those safeguards with an upcoming Claude Opus model [5]. Anthropic also says it will report back on what it learns from Glasswing [4].
- **U.S. officials are already in the loop on advanced cyber-capable AI:** Anthropic is in ongoing discussions with U.S. government officials about Mythos Preview and its offensive and defensive cyber capabilities [42].
- **Japan is already using AI for misinformation response:** Sakana AI says it completed a Ministry of Internal Affairs and Communications project to help visualize, assess, and propose countermeasures for large-scale online misinformation, and says it will continue contributing in intelligence-related AI work [43].

Quick Takes

Why it matters: Smaller releases still show where the market is heading: video generation, local fine-tuning, robotics, model serving, and increasingly fragmented product tiers.*

- **Dreamina Seedance 2.0** reached **#1** in Video Arena for both text-to-video and image-to-video [44].
- **DeepSeek** rolled out Fast/Expert/Vision-style chat modes, but early testers said Expert still looked closer to **V3.2** with about a **128K** context window than to a clearly new V4-class system [45, 46, 47].
- **Upstage Solar Pro 3** launched as a **102B MoE** model with **128K** context, strong instruction-following and tool-use scores, but a negative AA-Omniscience reliability result [48].
- **Gemma 4** is now available in the **Gemini API** and **Google AI Studio**, with support for function calling, image understanding, and Google Search grounding [49].
- **Unsloth** says Gemma 4 fine-tuning now works locally from **8GB VRAM**, with **1.5x** faster training and **50%** less VRAM use [50].
- **OpenAI Codex** will retire older models for ChatGPT sign-in users on **April 14** and move supported usage to newer GPT-5.4/5.3-era options [51, 52].

- **Weights & Biases + OpenPI** now support tracking physical-AI experiments, including fine-tuning robot foundation models on ALOHA with as little as **1 hour** of data [53].
- **GitHub Copilot** is now directly integrated into `code` itself [54].

Sources

1. X post by @AnthropicAI
2. X post by @AnthropicAI
3. X post by @AnthropicAI
4. X post by @AnthropicAI
5. X post by @AnthropicAI
6. X post by @sleepinyourhat
7. X post by @sleepinyourhat
8. X post by @sleepinyourhat
9. X post by @kimmonismus
10. X post by @Zai_org
11. X post by @Yuchenj_UW
12. X post by @OpenAIDevs
13. X post by @latentspacepod
14. X post by @sama
15. X post by @thsottiaux
16. X post by @JordiRib1
17. X post by @mustafasuleyman
18. X post by @mustafasuleyman
19. X post by @dejavucoder
20. X post by @algo_diver
21. X post by @bensig
22. X post by @nmboffi
23. X post by @ProfBuehlerMIT
24. X post by @cognition
25. X post by @cognition
26. X post by @cognition
27. X post by @cognition
28. X post by @kevinweil
29. X post by @kevinweil
30. X post by @kevinweil
31. X post by @awscloud
32. X post by @apsdehal
33. X post by @LangChain
34. X post by @hwchase17
35. X post by @kimmonismus
36. X post by @kimmonismus
37. X post by @dl_weekly

38. X post by @kimmonismus
39. X post by @kimmonismus
40. X post by @NousResearch
41. X post by @MiniMax__AI
42. X post by @scaling01
43. X post by @SakanaAILabs
44. X post by @arena
45. X post by @ZhihuFrontier
46. X post by @ZhihuFrontier
47. X post by @ZhihuFrontier
48. X post by @ArtificialAnlys
49. X post by @_philschmid
50. X post by @UnslothAI
51. X post by @OpenAIDevs
52. X post by @OpenAIDevs
53. X post by @wandb
54. X post by @pierceboggan