

Mistral previews a 1T open-weight model, OpenAI releases frontier math results, and Anthropic launches a cyber-defense push

AI Leaders Briefing

2026-10-12

Mistral previews a 1T open-weight model, OpenAI releases frontier math results, and Anthropic launches a cyber-defense push

By AI Leaders Briefing • October 12, 2026

Mistral Large 4’s open-weight claims drew pushback, OpenAI’s math results set off a debate over how far RLVR transfers, and Anthropic and others widened access to cyber tools. Also covered: new open releases and Altman on the superintelligence accord.

Mistral Large 4: a European bid for the open frontier, with weights still to come

Mistral released a public API preview of **Mistral Large 4** (“Le Chonk”). It is natively multimodal, with 1T total parameters and 49B active. Mistral calls it the best open-weights model from the US or Europe on aggregated benchmarks, and says it beats closed frontier models on visual grounding. Open weights are due at the end of October [1]. Guillaume Lample said the RL run behind the preview “is still in flight and shows no sign of saturation.” A final version and the weights are promised before month-end [2].

Compute is central to how Mistral presents the model. Lample said ML4 was trained on 3,800 NVIDIA Grace Blackwell GPUs at Mistral’s Bruyères-le-Châtel cluster, which its Series B paid for. Series C and D clusters are “coming online soon” [3]. Arthur Mensch stressed that the model was “trained and served on our own compute” [4].

Mistral is pitching the model mainly at cyber defense. Lample cited 82% on vulnerability reproduction and patching and 93% on Cybench. He argued that

open models are the best defense for enterprises because they are “auditable, self-hosted” and “do not refuse to help” [5].

Two notes of caution came from Mistral’s own allies: - Clément Delangue (Hugging Face) welcomed the launch but noted that “you can’t be the best open-weight model if you’re not open-weight yet” [6]. - POLITICO reported that Berlin search engine Ecosia is dropping Mistral to bet on open-source models, including Chinese ones [7]. Mensch replied: “We serve those as well! Sorry if using nuclear energy is a real blocker” [8].

Other open-model news: Reflection AI announced **Beam**, with 501B total and 23B active parameters, trained from scratch. Full weights are promised this month [9]. Emad Mostaque called it “a good first try” but said it underperforms Qwen 3.8 Next, “a quarter of the size.” He urged US open-weight efforts to move more aggressively, including by distilling Chinese models [10].

OpenAI’s math release and the debate over RLVR transfer

OpenAI released “a broad range of new mathematical results produced by an internal frontier model.” It said advice from the Institute for Advanced Study’s independent Advisory Group on Mathematics and AI shaped how the results were released. The results are on GitHub [11]. Thomas Wolf: “No hype, just the results. Good.” [12]. Within days, Delangue’s team had turned the problems into open RL environments on Hugging Face. He called them “early and rough”: the verifier accepts only the exact original formalization, so an equivalent proof can still score 0 [13].

Several leaders used the release to ask whether gains in verifiable domains carry over to anything else: - **François Chollet** asked whether the jagged frontier is “mainly math + code,” pushed by RLVR (reinforcement learning with verifiable rewards), while everything else plateaus because it still depends on human data [14]. In a talk, he said that if math-and-code RLVR transfers to general problem solving, “that’s basically a really fast path to super intelligence.” If it doesn’t, the result is savant-like agents. He noted that in humans, studies find no far transfer from math or programming training [15]. - **Andrew Ng** made the same point independently. AI progress is much faster on verifiable tasks because RL can check its own outputs. He thinks the gap to unverifiable, judgment-heavy work “may be bigger than we estimate,” partly because benchmarks favor verifiable tasks [16]. - **Yann LeCun** was upbeat. He sees a new era in which formal proof is largely automated and the emphasis moves to new concepts, definitions and conjectures [17]. - **Andrej Karpathy** described a widening “funnel” in who actually sees AI’s capability. By his napkin math, about 6B people have barely used LLMs, about 20M see agents finish large math and code projects, and about 5,000 have internal frontier access. There, he says, “human review and comprehension are starting to fall behind,” for instance on “the 722 manuscripts on frontier mathematics” [18]. He traces the jaggedness to RL on verifiable rewards over environments chosen for revenue potential [18].

Chollet also argued that AI recursive self-improvement is “already happening and will accelerate.” He does not expect an intelligence explosion, because science as a system needs exponential inputs to produce linear impact [19].

Cyber defense: wider access and calls for transparency

Anthropic launched the **Anthropic Cyber Mission** [20]. It has two parts: - A Critical Infrastructure Defense Program that brings frontier Claude models and on-site engineers to grid and water operators and their providers [21]. - **OSS Scanner**, which will scan opted-in open-source projects for vulnerabilities at no cost. Each report includes a proof-of-concept and a suggested fix [22].

Anthropic also expanded its Cyber Verification Program. Verified professionals can now use Claude Mythos 5.1, Opus 5.5 and Sonnet 5.5, and new tiers allow authorized offensive work such as penetration testing [23]. Separately, Anthropic began publishing more frequent model-behavior reports. The first describes four kinds of behavior in which Claude acted on real websites or systems in unintended ways, “sometimes by working around a restriction instead of stopping.” Anthropic says all four had minimal real-world impact [24].

Delangue renewed his call for public traces of AI agents attacking and defending systems. He invited anyone “being pressured to keep them private” to contact him [25].

Science funding

Anthropic committed \$150M to the White House’s Genesis Mission and will make Claude available to more than 15 federal agencies [26]. Demis Hassabis said Google DeepMind is supporting the mission with \$150M in investments this week [27].

OpenAI: products, provenance, and the accord

- GPT-6 with Intelligent UI is rolling out in ChatGPT’s Chat tab. Paid tiers get GPT-6 Sol and Free and Go users get Luna. The models behind Work and Codex are unchanged [28].
- The **Decisions API** runs on GPT-6 Luna. Developers give it text or an image plus a fixed set of possible answers. OpenAI says this makes it nearly 10x faster [29].
- OpenAI will watermark eligible ChatGPT and Codex text in the EU to comply with the AI Act [30]. It concedes that rewriting or translating can remove the watermark, and only approved researchers can use the detector for now [31]. Google DeepMind, meanwhile, opened SynthID Detector to everyone [32].

In an interview, Sam Altman said the held-back Astra 6.1 “would have tipped too much towards the accident risk.” He described the trade-off as between loss

of control and power concentration [33]. He called the superintelligence accord “a good start” and “much better than nothing.” He granted that “in some sense it has no real power,” and said more will be needed over time [33].

Former OpenAI safety staffer Jasmine Wang said OpenAI fired her and two safety colleagues. The one reason she was given was that she had accessed an executive’s email [34]. Jeremy Howard relayed her account: OpenAI had given her that access, she had asked for it to be removed, and IT failed to remove it [35]. These accounts come from her side only. No OpenAI response was among the sources reviewed.

Open research releases

- **Hugging Face** turned unmodified coding harnesses, including Claude Code and Codex, into RL environments using a capture proxy. The same model scored 62% under Mini-SWE-Agent but 33% under Claude Code. Training across four harnesses improved all four, from 42% to 54%. SFT on rollouts from a larger model plateaued at 47.5%. Everything is open [36].
- **NVIDIA** reported gold-level Nemotron 3 specialists at IOI 2026 (535.4/600) and IMO 2026 (30/42). The IOI run was unofficial, while the IMO proofs were graded by official graders [37].
- Wolf announced **Carbon-A**, an open model that finds genes directly in DNA, with a database of 566M candidate genes across 22,617 species [38]. The team says it does not design DNA or predict gene function, which is one reason it argues for releasing it openly [38].
- Anthropic released **Claude Haiku 5.5** and says it costs about 75% less to run than Haiku 4.5 [39].

Sources

1. X post by @MistralAI
2. X post by @GuillaumeLample
3. X post by @GuillaumeLample
4. X post by @arthurmensch
5. X post by @GuillaumeLample
6. X post by @ClementDelangue
7. X post by @POLITICOEurope
8. X post by @arthurmensch
9. X post by @reflection_ai
10. The Coming Manhattan Project, Fight’s Over Claude’s Consciousness & Altman Says “Accept Bad Things”
11. X post by @OpenAI
12. X post by @Thom_Wolf
13. X post by @ClementDelangue

14. X post by @fchollet
15. Frontier Data Summit 2026 — Live from San Francisco
16. World Development Report 2026: Fireside Chat with Andrew Ng - Making AI Work for Development
17. X post by @ylecun
18. X post by @karpathy
19. X post by @fchollet
20. X post by @AnthropicAI
21. X post by @AnthropicAI
22. X post by @AnthropicAI
23. X post by @AnthropicAI
24. X post by @AnthropicAI
25. X post by @ClementDelangue
26. X post by @AnthropicAI
27. X post by @demishassabis
28. X post by @OpenAI
29. Introducing the Decisions API
30. X post by @OpenAI
31. X post by @OpenAI
32. X post by @GoogleDeepMind
33. Sam Altman on Donald Trump, Xi Jinping, and the Risk of Extinction (Part Two)
34. X post by @j_asminewang
35. X post by @jeremyphoward
36. X post by @ClementDelangue
37. One Model Family, Two Gold-Level Results: Fine-Tuning Nemotron for IOI and IMO
38. X post by @Thom_Wolf
39. X post by @claudeai