

# Gemini 4 Argon goes global as Meta and Reflection release US open-weight frontier models

AI Leaders Briefing

2026-10-19

## Gemini 4 Argon goes global as Meta and Reflection release US open-weight frontier models

*By AI Leaders Briefing • October 19, 2026*

Google DeepMind made Gemini 4 Argon generally available with strong ARC-AGI results and science demos. Meta’s Muse Spark Open and Reflection’s Beam shipped open weights. Jack Clark told the IMF the singularity has begun, and LeCun, Ng and Chollet pushed back.

### Gemini 4 Argon: from trusted testers to general availability

Last brief, Argon was only with trusted testers. This week Sundar Pichai made it available worldwide in the paid Gemini API and the Gemini platform. Ultra subscribers get it in Antigravity, and Pro and Ultra subscribers get it in the Gemini app and Search [1].

ARC Prize verified these scores [2]: - **ARC-AGI-3:** 97.5% for \$4.8K with the provider’s adapter harness, but 49.8% for \$5.8K with the standard harness. - **ARC-AGI-2:** 91.3%, at \$0.32 per task. - Google DeepMind is the second frontier lab to score above 95% on ARC-AGI-3.

François Chollet said Argon “has a similar performance profile on ARC-AGI-3 as Astra” and that “Google is back at the frontier” [3]. The large gap between the two harnesses is worth keeping in mind.

DeepMind researchers posted several science results: - **Quantum computing:** Argon translated an internal version of the AlphaQubit decoder into a single hardware “megakernel,” nearly doubling inference throughput [4]. - **Plasma physics:** Opal Issan reported two new families of counterexamples to the nearly 60-year-old Grad’s conjecture, in a paper with University of Maryland collaborators [5]. - **Enzymes:** Argon designed five lab-verified serine esterases in a matter of hours [6].

DeepMind is offering 10,000 scientists six months of free access to Antigravity’s highest usage tier [7].

In a separate test, four general-purpose Argon coding agents defeated Minecraft’s Ender Dragon. They worked through the bare Mineflayer library, with no hints, assigned roles or human steering [8]. The server log records the win 27 hours 14 minutes after the first program, across 2,783 programs [8]. The team listed its own caveats: the agents read the world through an API, used a pathfinding plugin, played on Easy, respawned 605 times, and 22% of their programs ended in an error [8].

## Open weights at US frontier scale and Europe’s case for building

Meta released **Muse Spark Open** and called it “the most capable US open-weight model to date.” It has 816B total parameters (42B active), a 1M-token context, and is natively multimodal [9]. Every one of its 256 experts must stay in memory, so the BF16 checkpoint is about 1.5 TB. It runs on four 8×H200 nodes, or on a single B200 node in NVFP4 [10]. Mark Zuckerberg said it “runs great on new Mac Studios” after work with Apple [11].

Reflection has now actually shipped **Beam**, which was only promised last period. It is a 501B-parameter MoE (23B active), text-only, with a 1M-token context, released under Apache 2.0 and aimed at coding and agents [12].

DeepSeek called V4.1-Flash-1017 a “minor update,” but Thomas Wolf noted it moved Terminal-Bench 4.0 from 31.2 to 42.3. He also said the more interesting claim is shorter outputs. He wants to see whether the numbers hold up outside DeepSeek’s own evaluation setup [13].

Arthur Mensch’s essay “Build or surrender” argues that Europe cannot leave frontier models to the US and China. His argument: “whoever controls the model sets the terms.” Even relying on others’ open models is still dependence, because you cannot choose what gets released, when, or under what license [14]. He said Mistral will increase its compute capacity sixfold over the coming year and will raise more capital [14].

Wolf added data. FLUX.1-dev from Germany’s Black Forest Labs is the Hub’s most built-upon model, with 43k derivatives. Europe trained about 8% of the Hub’s open-model compute, yet about 18% of what is built on top of it is European [15]. He estimates roughly \$1B of training compute is now shared on the Hub, up from under \$40M at the end of 2023. He calls this a floor, about half of it an educated guess [16].

Mistral also released **Timestral**, a family of time-series models. Mistral says its largest, at 209M parameters, ranks first on GIFT-Eval, TIME and fev-bench [17].

## Has the singularity begun? Clark says yes, others push back

In the IMF Per Jacobsson Lecture, Jack Clark said “the singularity is here” [18]. His evidence: - Anthropic researchers now produce almost 10× as much code as in the company’s first four years [18]. - Humanity’s Last Exam scores rose from 8% to 67.7%, and Frontier Math from 2% to 94% [18]. - On an unpublished theoretical-physics test, GPT-6 Astra reached 49/50 after experts found errors in their own answers [18].

He paired this with a call for “restraint” and to “go slowly and deliberately” on developing the technology [18]. Earlier in the week, he put “a better than even chance” on a system being able to build its own successor in 2028 [19].

Critics pushed back: - **Yann LeCun**: “It really has not... Unless you get fooled into thinking that every point on an exponential curve is a singularity” [20]. - **Andrew Ng** called recursive self-improvement “much more hype than reality.” He said progress is held back by data centers, power plants and experiments [21]. He also said no software engineer who can’t use a coding agent gets hired on his teams [21]. - **Chollet**: engineers now rarely write code themselves, since agents write about 100% of it, yet employers’ bottom lines are “typically not budging.” His point is that automating one part of a system moves the bottleneck elsewhere [22]. He also said Astra is better than him at coding but cannot grasp simple aspects of his research [23]. - **Demis Hassabis** described this year’s hacking incidents as “early warning shots.” He called for risk-proportionate certification that can change quickly as the technology moves [24].

## Anthropic: equity for the public good, outside evaluators, agent-scale engineering

Anthropic created a **Public Benefit Fund** that starts with 5% of its shares. It can receive up to 2% more for every 50% rise in company value over a set baseline, at the board’s discretion. Anthropic says this brings roughly 20% of its shares committed to the public good [25]. Clark said the money is meant to help more people share in AI’s benefits and to fund work on its risks [26].

Anthropic has also given a METR team employee-like access to assess its safety practices, and is in talks with other evaluators [27]. Clément Delangue said open-source scientists should be included in such efforts. He noted that a Hugging Face role “hasn’t been approved yet” and that Wolf is starting an open alignment initiative [28].

On engineering, Anthropic’s daily merged code changes went from about 2,500 in May to more than 30,000 in September. That outgrew git, so it built its own version-control system and moved its monorepo onto it in five months [29]. Separately, its STEM Fellows report that Claude Mythos 5 designed quantum circuits that beat the best published ones, sometimes by an order of magnitude or more [30].

OpenAI and Anthropic each committed \$5M to Simons Foundation initiatives that help mathematicians examine and build on AI-produced results [31, 32].

## Cohere moves on cyber defense and sovereignty

Aidan Gomez said Cohere is bringing in Aisle; the linked post describes a letter of intent. He says Aisle has found more than 200 High or Critical CVEs in software such as curl and FFmpeg [33]. He also claimed Aisle found 29 new curl CVEs where Mythos found zero, citing curl’s maintainer [34]. Cohere plans to offer the scanning to critical-infrastructure operators, including fully air-gapped deployments [35].

In an interview, Gomez defined sovereignty as diversifying suppliers so that “no one party can shut you off,” not every country building its own model [36]. He pointed to the US government cutting off non-American access to Mythos, which the interviewer described [36], and called it “crossing the Rubicon” [36].

## NVIDIA

NVIDIA released **Nemotron 3.5 Super VL**, an open model for long-running agents that understands text, images, documents and video. It has 120B total parameters, 12B active [37].

Artificial Analysis published first results for Vera Rubin NVL72, all submitted by NVIDIA. It shows up to about  $5.7\times$  GB300’s throughput per megawatt on DeepSeek V4 Pro [38]. The advantage is much smaller at peak throughput, at  $1.14\text{--}1.45\times$  [38].

---

## Sources

1. X post by @sundarpichai
2. X post by @arcprize
3. X post by @fchollet
4. X post by @benjaderberg
5. X post by @opalissan
6. X post by @GoogleDeepMind
7. X post by @GoogleDeepMind
8. X article by @maxencefaldor
9. X post by @AIatMeta
10. Welcome Muse Spark Open: open-weight frontier multimodal mixture-of-experts model from Meta
11. X post by @finkd
12. Welcome Beam: a 501B open model you can own, customize, and scale
13. X post by @Thom\_Wolf
14. X article by @arthurmensch
15. X post by @Thom\_Wolf

16. X post by @Thom\_Wolf
17. Introducing Timestral
18. Per Jacobsson Foundation Lecture: Inside the Event Horizon
19. Anthropic's Jack Clark on AI, power, and the Pentagon
20. X post by @ylecun
21. AI won't steal your job: Andrew Ng interview | The Economist
22. X post by @fchollet
23. ARC-AGI Explained: Closing the Human-AI Gap — François Chollet
24. BREAKING: Google DeepMind Chief Warns AI Hacks Are “Early Warning Shots” of Bigger Threats | AI1E
25. X post by @AnthropicAI
26. X post by @jackclarkSF
27. X post by @AnthropicAI
28. X post by @ClementDelangue
29. X post by @AnthropicAI
30. X post by @AnthropicAI
31. X post by @OpenAI
32. X post by @AnthropicAI
33. X post by @aidangomez
34. X post by @aidangomez
35. X post by @aidangomez
36. Aidan Gomez, Co-Founder and CEO of Cohere [AI Founders Masterclass 2026 Session 3]
37. X post by @NVIDIAAI
38. X post by @ArtificialAnlys