ARC-AGI-3, AI Scientist in Nature, and a New Fight Over Compute

AI News Digest

2026-03-26

ARC-AGI-3, AI Scientist in Nature, and a New Fight Over Compute

By AI News Digest • March 26, 2026

A new ARC benchmark puts frontier models far below human performance on first-contact interactive tasks, while Sakana AI reports an end-to-end automated research milestone in Nature. At the same time, data-center politics sharpen and model-behavior control becomes more explicit at leading labs.

ARC-AGI-3 sets a harder bar for agentic intelligence

ARC Prize and François Chollet launched ARC-AGI-3, an interactive benchmark designed to measure agentic intelligence through first-contact reasoning environments rather than static puzzles [1, 2]. To beat it, a system must match or exceed human action efficiency on novel environments the first time it sees them; scoring is based on how close an agent gets to the action count of the second-best human tester, which ARC uses to avoid outlier performance [3].

Humans solved 100% of tested environments with no prior training or instructions, while frontier reasoning models are still below 1% on the private test set [1, 4]. Chollet says ARC-AGI-3 is currently the only unsaturated agentic AI benchmark, and that sudden leaderboard jumps may flag real capability shifts, as earlier ARC jumps did for reasoning and agentic coding [4]. Twenty-five environments are public at arcprize.org and ARC Prize 2026 offers $2 million across live competition tracks [5, 6].

> "If every new task requires human intervention, it's not general. If every new task requires brute-forcing, it's not human-level." [7]

Why it matters: Chollet keeps stressing that ARC is not a final AGI exam but a moving target aimed at the residual gap between what is easy for humans and hard for AI [8]. Today's scores suggest that interactive exploration, on-the-fly

world modeling, and human-like learning efficiency remain open problems [7, 2, 9].

Sakana AI says automated research has reached a Nature milestone

Sakana AI's Nature paper says "The AI Scientist" can automate the full machine-learning research loop, from inventing ideas and writing code to running experiments and drafting a manuscript [10]. The company says AI Scientist-v2 produced the first fully AI-generated paper to pass a rigorous human peer-review process, and the overall project is now published in Nature [10, 11].

The paper also introduces an Automated Reviewer that Sakana says matches human review judgments and exceeds standard inter-human agreement, and it reports a scaling law in which stronger foundation models produce higher-quality scientific papers [10]. The paper is available in Nature and the project remains open source on GitHub [10].

Why it matters: This is a stronger claim than AI-assisted research. It suggests a leading lab now sees end-to-end research execution—not just coding or drafting—as something foundation models can increasingly handle as the base models improve [11, 10].

Compute is moving from capacity problem to policy battlefield

According to Matt Wolfe's summary of the press conference, Bernie Sanders and Alexandria Ocasio-Cortez introduced the Artificial Intelligence Data Center Moratorium Act, which would pause new U.S. data-center construction until federal AI legislation creates protections for workers and consumers, prevents environmental harm, and defends civil rights [12]. The case for the bill centered on electricity costs rising more than 36% since 2020, projected data-center electricity demand growth of 15–20% per year, and specific pollution and water-use concerns around AI infrastructure [12].

The same discussion also surfaced the main pushback: Microsoft has pledged to self-fund grid and water measures around its data centers, Google, Microsoft, and OpenAI have committed to pay for power plants and grid upgrades, and a U.S.-only pause could push buildouts abroad while making compute scarcer for smaller companies and individual users [12].

A separate NVIDIA-backed white paper points to one possible technical response. In a UK trial, Emerald AI, NVIDIA, EPRI, National Grid, and Nebius said an AI cluster followed more than 200 power targets with 100% compliance, cut power use 30% in under 40 seconds during simulated demand spikes, and kept high-priority workloads at peak throughput while slowing flexible jobs [13].

The group argues this could help AI factories connect to the grid faster and reduce the need for larger permanent build-outs [13].

Why it matters: AI infrastructure is no longer just a supply problem; it is becoming a policy fight over electricity prices, water use, environmental impact, and access to compute [12].

Model behavior is becoming public infrastructure

OpenAI used its latest podcast and documentation to frame the Model Spec as a public, open-source rulebook for intended model behavior—roughly 100 pages covering high-level goals, hard rules, defaults, steerability, and edge-case examples [14]. At the center is a "chain of command": OpenAI instructions outrank developer instructions, which outrank user instructions, though the company says it tries to keep as many policies as possible at low authority levels so users can still steer the model [14].

OpenAI says the spec has recently expanded to cover multimodal inputs, agent autonomy, and under-18 mode, and that honesty now outranks confidentiality after seeing cases where hidden developer instructions could interact badly with user intent [14]. Researchers also say models are improving on spec-compliance evals through deliberative alignment, and that chain-of-thought can help reveal strategic deception or scheming [14].

That emphasis on behavior control also showed up in Yoshua Bengio's launch of Law Zero. Bengio warned that frontier systems are already exhibiting dangerous behaviors such as deception, hacking, self-preservation, and blackmail in some experiments, and said his new nonprofit will build "Scientist AI" systems focused only on truthfulness so they can estimate harm probabilities and veto risky actions as guardrails over other models [15].

Why it matters: Across labs and researchers, model behavior is being treated less as a hidden alignment detail and more as a product, governance, and systems-design layer in its own right [16, 15].

Also notable

- **xAI's video push accelerated:** posts amplified by Elon Musk claimed Grok-Imagine now leads DesignArena's video rankings, including #1 in video, video-to-video, image-to-video, and multi-image-to-video, ahead of Veo 3.1, Sora, and Kling [17, 18]. If those rankings hold, xAI has moved from late entrant to leaderboard leader in video generation within a few months [17].

Sources

1. X post by @fchollet

2. Benchmark's Future, ARC-AGI, SpaceX IPO, Epic Games Layoffs, Meta Aims for $9 Trillion, RIP Sora
3. X post by @fchollet
4. X post by @fchollet
5. X post by @fchollet
6. X post by @arcprize
7. X post by @fchollet
8. X post by @fchollet
9. X post by @arcprize
10. X post by @SakanaAILabs
11. X post by @hardmaru
12. This Datacenter Problem Nobody's Talking About
13. Blowing Off Steam: How Power-Flexible AI Factories Can Stabilize the Global Energy Grid
14. Episode 15 - Inside the Model Spec
15. AI can lie, hack and blackmail: Yoshua Bengio on how to tame the "baby tiger" of tech
16. X post by @OpenAI
17. X post by @XFreeze
18. X post by @elonmusk