AI Scientist Reaches Nature as ARC-AGI-3 Debuts and GPT-5.4 Gets Cheaper

AI High Signal Digest

2026-03-26

AI Scientist Reaches Nature as ARC-AGI-3 Debuts and GPT-5.4 Gets Cheaper

By AI High Signal Digest • March 26, 2026

Sakana AI's Nature paper, ARC-AGI-3's human-AI gap, and OpenAI's GPT-5.4 mini and nano headline the cycle. The brief also covers new research architectures, product rollouts, hiring and funding signals, and the latest policy and governance moves.

Top Stories

Why it matters: This cycle mixed a research milestone, a new benchmark gap, cheaper frontier-model variants, and a deployment-level inference breakthrough.

Sakana AI took *The AI Scientist* into *Nature*

Sakana AI said *The AI Scientist: Towards Fully Automated AI Research* is now published in *Nature* [1]. The system is described as an agent built from foundation models that can run the full machine-learning research loop: invent ideas, write code, run experiments, and draft the paper [1]. Sakana also said AI Scientist-v2 produced the first fully AI-generated paper to pass rigorous human peer review, and that the *Nature* paper introduces an Automated Reviewer that matches human judgments and exceeds standard inter-human agreement [1]. The paper reports a "scaling law of science": stronger foundation models—and, in later commentary, more inference compute—produce higher-quality generated papers [1, 2]. The work is open-source and was done with collaborators at UBC, the Vector Institute, and Oxford [1].

Why it matters: this is one of the clearest public attempts to combine end-to-end research automation, peer-reviewed validation, and open release in a single result.

ARC-AGI-3 opened with a wide human-AI gap—and immediate debate about the metric

ARC-AGI-3 was released as a benchmark for agentic intelligence in interactive reasoning environments, with the stated goal of measuring whether an AI can match human-level action efficiency on unseen tasks [3]. ARC Prize said humans solve 100% of environments on first contact with no prior training or instructions, while frontier AI models are under 1% at launch [3]. A set of posted scores put Gemini 3.1 Pro at 0.37%, GPT-5.4 at 0.26%, Opus 4.6 at 0.25%, and Grok 4.2 at 0% [4]. François Chollet separately said ARC-AGI is *not* a final exam for AGI, but a moving target aimed at the residual gap between what is easy for humans and hard for AI [5, 6].

> "Most benchmarks test what models already know, ARC-AGI-3 tests how they learn" [7]

The benchmark design is already under scrutiny. Official posts say the human baseline uses the action count of the second-best tester out of 10, and a score measures how close a system gets to matching or exceeding that baseline [8]. External commentary noted quadratic scaling of steps and warned that ARC-AGI-3 scores should be interpreted differently from standard benchmarks [9], while other critics questioned the "human score 100%" framing and whether prior puzzle or game exposure makes the human comparison less clean than advertised [10, 11].

Why it matters: ARC-AGI-3 is now both a hard new public target for agentic systems and a live debate over how progress should be measured.

OpenAI widened the GPT-5.4 line with cheaper mini and nano models

Artificial Analysis reported that OpenAI released GPT-5.4 mini and GPT-5.4 nano, both with the same reasoning effort modes as GPT-5.4, multimodal image input, and a 400K-token context window [12]. Pricing was listed at $0.75/$4.50 per 1M input/output tokens for mini and $0.20/$1.25 for nano, versus $2.50/$15 for GPT-5.4 [12]. The same evaluation said nano outperformed Claude Haiku 4.5 and Gemini 3.1 Flash-Lite Preview on several reasoning and terminal-style tests, while mini posted stronger agentic GDPval-AA scores than Gemini 3 Flash Preview but trailed Claude Sonnet 4.6 [12, 13]. The tradeoff is efficiency: both models used far more output tokens than peers at highest reasoning effort, and both showed weak AA-Omniscience results driven by high hallucination rates [12, 14].

Why it matters: OpenAI is pushing its frontier line further downmarket, but the benchmark data suggests buyers still need to watch token consumption and hallucination behavior.

TurboQuant moved from paper result to open inference deployment

Google Research introduced TurboQuant as a compression algorithm that cuts LLM key-value cache memory—the working memory models use during generation—by at least 6x and delivers up to 8x speedup with zero accuracy loss [15]. A separate technical summary said the method needs no retraining, converts data into polar coordinates to remove storage overhead, and applies a 1-bit correction step; tests on Gemma and Mistral models reportedly matched full-precision quality on question answering and code generation while also beating prior methods in vector search [16]. The result quickly showed up in the open serving stack: one developer said they implemented TurboQuant for vLLM and fit 4,083,072 KV-cache tokens on a USB-charger-sized HP ZGX, which the vLLM project then praised publicly [17, 18].

Why it matters: this is a case where an inference paper is already showing concrete deployment effects in open tooling.

Research & Innovation

Why it matters: Beyond the headline stories, this cycle emphasized self-improving agents, shared memory, hybrid architectures, and native multimodality.

- **Hyperagents:** Meta and collaborators introduced self-referential agents where the self-improvement process itself is editable, rather than fixed [19]. The DGM-Hyperagent combines a task agent and a meta agent in one modifiable program, discovering improvements such as persistent memory and performance tracking that transfer across domains [19]. Reported gains included paper review accuracy moving from 0.0 to 0.710, robotics reward design from 0.060 to 0.372, and zero-shot transfer to Olympiad-level math grading at 0.630 [19].
- **MemCollab:** New research on memory sharing across heterogeneous agents uses contrastive trajectory distillation to separate universal task knowledge from agent-specific biases [20]. In plain terms, it compares how different agents reason through the same task to extract shared constraints, then uses task-aware retrieval to apply the right constraints later [20]. The authors report gains in both accuracy and inference-time efficiency for math reasoning and code generation, even across model families [20].
- **Hybrid Associative Memory (HAM):** ZyphraAI proposed a Transformer/RNN hybrid that lets the RNN handle predictable tokens and the Transformer handle surprising ones based on a user-selected KV-cache budget [21, 22]. At 800M parameters, HAM was reported to outperform pure Transformer, pure RNN, and prior hybrid baselines on language modeling and long-context retrieval while using only 50% KV cache [23]. The architecture also allows adjustable KV cache at inference time and even within a single sequence [24].

- **LongCat-Next:** Meituan introduced a native autoregressive multimodal model with 68.5B total parameters and 3B active parameters, built on a shared discrete token space across language, vision, and audio [25]. The model combines a new any-resolution vision transformer with capabilities in OCR, charts, GUI understanding, document analysis, arbitrary-resolution visual generation, audio comprehension, and voice cloning [25].

Products & Launches

Why it matters: New releases this cycle were less about one giant model launch and more about turning AI into usable, task-specific software.

- **AssemblyAI Medical Mode:** AssemblyAI added a medical correction layer on top of Universal-3 Pro, aimed at fixing the drug names, dosages, and terminology errors that make general-purpose ASR unsafe for clinical workflows [26]. The company says the base model's noise handling and latency stay the same, while the correction focuses on key medical tokens; it is available for both pre-recorded and streaming audio, with HIPAA BAA included [26, 27].
- **Lyria 3 Pro rollout:** Google DeepMind and Gemini said Lyria 3 Pro now supports tracks up to three minutes, with structure controls for intros, verses, choruses, and bridges [28, 29]. Access is rolling out in the Gemini App for Google AI Plus, Pro, and Ultra users, while developers can build against it in Google AI Studio and the Gemini API [29, 30, 31]. Google also said all Lyria 3 and Lyria 3 Pro outputs carry SynthID watermarking [32].
- **Claude work tools on mobile:** Anthropic said Claude's work tools are now available on mobile, including access to Figma designs, Canva slides, and Amplitude dashboards from a phone [33].
- **Cursor self-hosted cloud agents:** Cursor said its cloud agents can now run on customer infrastructure, keeping code and tool execution inside the user's own network while preserving the same agent harness and experience [34].
- **LangSmith Fleet shareable skills:** LangChain added shareable skills to LangSmith Fleet, letting teams capture domain knowledge once, attach it to any agent, and create skills from prompts, past chats, manual entry, or templates [35, 36].

Industry Moves

Why it matters: Hiring patterns, partnerships, and funding are showing where companies think the next wave of value will come from.

- **AI labs are hiring for go-to-market and adoption at scale:** Epoch AI's analysis of job postings at OpenAI, Anthropic, xAI, and DeepMind said sales and go-to-market roles are now the largest hiring category at OpenAI and Anthropic, at 31% and 28% of open roles respectively, while research roles account for 7% and 12% [37, 38]. The same analysis pointed to heavy hiring for “AI Success Engineer” and “Forward Deployed Engineer” roles, 15 OpenAI roles tied to a consumer hardware device, and growing investment in robotics at both OpenAI and DeepMind [39, 40].

- **Cohere partnered with RWS:** Cohere said its frontier models are being integrated into RWS Group’s Language Weaver Pro to provide enterprise-grade translation for high-stakes environments, including enterprise and government use cases [41].
- **Gumloop raised $50M:** Gumloop raised a $50M Series B led by Benchmark, bringing total funding to $70M for its no-code AI agent automation platform [42].
- **AirStreet closed a larger AI-first fund:** AirStreet said it raised $232,323,232 for Fund III to back AI-first companies in the U.S. and Europe, making it the largest solo GP venture firm in Europe by its own description [43].

Policy & Regulation

Why it matters: AI policy is now reaching physical infrastructure, while labs are continuing to publish formal governance frameworks for model behavior.

- **Sanders targets data-center buildout:** The *Washington Post* said Sen. Bernie Sanders will introduce legislation to block construction of new data centers until lawmakers enact AI regulations [44].
- **OpenAI highlighted its Model Spec:** OpenAI described the Model Spec as the public framework for how its models are intended to behave, covering what they should and should not do as capability grows [45]. The company said the framework includes a chain of command for resolving conflicting instructions and evolves over time through real-world use, feedback, and new model capabilities [45, 46].
- **Anthropic documented auto-mode safety decisions:** Anthropic said Claude Code auto mode is meant to be a safer middle ground between prompting for approval on every action and running without permission prompts, using built and tested classifiers to make approval decisions [47].

Quick Takes

Why it matters: These items were smaller, but they point to where tooling, interfaces, and agent infrastructure are moving next.

- Google Research’s **Vibe Coding XR** turns prompts into interactive, physics-aware WebXR apps through Gemini Canvas and XR Blocks [48]
- **LLaDA2** became the first discrete diffusion pipeline for text in Diffusers; it uses a 16B total-parameter MoE architecture [49, 50]
- **Browserbase** and **PrimeIntellect** launched BrowserEnv so users can train browser agents or custom models for their own workflows in a few hours [51, 52]

- A **24B** model was shown running locally in a web browser at about **50 tokens/sec** on an M4 Max using WebGPU and Transformers.js [53]
- Georgia Tech SSLab's **Vibe Radar** tracks public CVEs linked to AI-generated code, scanning 50k+ advisories and finding dozens of confirmed cases across tools such as Claude Code, Copilot, and Cursor [54]
- Anthropic launched **inline interactive charts, diagrams, and visualizations** in Claude chat, in beta across all plan types [55]
- **Together AI** added four new image models spanning text rendering, character consistency, search-grounded generation, and unified generation/editing on its serverless stack [56, 57, 58, 59, 60]
- **ARC Prize 2026** went live with three tracks and $2,000,000 in prizes [61]

Sources

1. X post by @SakanaAILabs
2. X post by @SakanaAILabs
3. X post by @fchollet
4. X post by @scaling01
5. X post by @fchollet
6. X post by @fchollet
7. X post by @arcprize
8. X post by @fchollet
9. X post by @FakePsyho
10. X post by @DeryaTR_
11. X post by @JJitsev
12. X post by @ArtificialAnlys
13. X post by @ArtificialAnlys
14. X post by @ArtificialAnlys
15. X post by @GoogleResearch
16. X post by @kimmonismus
17. X post by @iotcoi
18. X post by @vllm_project
19. X post by @omarsar0
20. X post by @dair_ai
21. X post by @ZyphraAI
22. X post by @ZyphraAI
23. X post by @ZyphraAI
24. X post by @ZyphraAI
25. X post by @Meituan_LongCat
26. X post by @AssemblyAI
27. X post by @AssemblyAI
28. X post by @GoogleDeepMind
29. X post by @GeminiApp

30. X post by @GeminiApp
31. X post by @Google
32. X post by @GeminiApp
33. X post by @claudeai
34. X post by @cursor_ai
35. X post by @LangChain
36. X post by @hwchase17
37. X post by @EpochAIResearch
38. X post by @EpochAIResearch
39. X post by @EpochAIResearch
40. X post by @EpochAIResearch
41. X post by @cohere
42. X post by @dl_weekly
43. X post by @nathanbenaich
44. X post by @washingtonpost
45. X post by @OpenAI
46. X post by @OpenAI
47. X post by @AnthropicAI
48. X post by @GoogleResearch
49. X post by @RisingSayak
50. X post by @RisingSayak
51. X post by @browserbase
52. X post by @johannes_hage
53. X post by @xenovacom
54. X post by @hankein95
55. X post by @dl_weekly
56. X post by @togethercompute
57. X post by @togethercompute
58. X post by @togethercompute
59. X post by @togethercompute
60. X post by @togethercompute
61. X post by @arcprize