Anthropic's Self-Improvement Metrics, Nemotron 3 Ultra, and Live Agent Evals

AI High Signal Digest

2026-06-05

Anthropic's Self-Improvement Metrics, Nemotron 3 Ultra, and Live Agent Evals

By AI High Signal Digest • June 5, 2026

Anthropic published unusually concrete data on AI-assisted AI development, NVIDIA released a major open agent model, and Agent Arena introduced a live benchmark for real-world agent performance. The brief also covers ChatGPT memory, enterprise retrieval, outcome-based AI go-to-market moves, and new policy attention on biosecurity and national AI strategy.

Top Stories

Why it matters: today's biggest developments were about AI improving AI, stronger open models, and better measurement of real agent performance.

- **Anthropic put hard numbers on AI-assisted AI development.** Anthropic said internal data shows Claude is accelerating AI development, with engineers shipping **8x** more code, Claude writing **80%+** of merged code, open-ended task success reaching **76%**, and the length of tasks AI can reliably complete doubling roughly every **4 months**. Anthropic outlined three futures—stalling progress, compounding gains with humans still setting direction, or full recursive self-improvement—and said the middle path is the likeliest. OpenAI separately said it also sees early signs of recursive self-improvement and warned existing institutions are not ready for the governance challenges. [1, 2, 3, 4, 5]
- **NVIDIA raised the bar for open agent models with Nemotron 3 Ultra.** The new model is a fully open **550B** model with **55B active parameters**, designed for long-running agents, up to **1M** context, and released with weights, training data, and recipe. NVIDIA says it delivers **5x** faster inference and up to **30%** lower cost on complex agentic tasks; Artificial Analysis said it now leads U.S. open-weight models on its Intelligence

Index at **47.7**. [6, 7, 8, 9]

- **Agent Arena launched a live benchmark for real agent work.** Arena said its new leaderboard is built from **300K+** tasks, **2M+** tool calls, and **40M** lines of code across live user sessions using web search, filesystem, and terminal tools. The first ranking places **OpenAI GPT-5.5** first, **Anthropic Claude-Opus-4.7** second, and **Z.ai GLM-5.1** third, signaling a shift away from static agent evals toward production-like measurements. [10]

Research & Innovation

Why it matters: the most useful research updates focused on long-horizon agents, multimodal grounding, and model oversight.

- **AutoLab argued that persistence matters more than first-try quality.** Across **17** frontier models and **36** expert-curated tasks in optimization, model development, CUDA kernels, and puzzles, the strongest predictor of success was repeated benchmarking, editing, and feedback loops—not the initial answer. The authors said Claude-opus-4.6 sustained that loop best. [11]
- **AllenAI's Molmo2 pushed open video-grounded vision forward.** The model supports video pointing, tracking, counting by pointing, and multi-image reasoning in one open system, returns precise pixel coordinates and timestamps, and was trained on new video and multi-image datasets collected without distilling from closed models. [12, 13, 14]
- **Goodfire showed a cheaper way to detect eval awareness.** Its new method uses logits to measure how close a model is to recognizing that it is being tested, reportedly requiring **10x to 100x fewer samples** than monitoring outputs alone. [15]

Products & Launches

Why it matters: consumer and enterprise AI products kept moving toward better memory, faster retrieval, and bigger working context.

- **OpenAI rolled out a more capable ChatGPT memory system.** The update carries context across conversations, lets users review and steer memory through a summary, and gives Plus and Pro users in the U.S. **2x** more memory. Team posts said the work evolved from saved memory to dreaming and now dreaming V3. [16, 17, 18, 19]
- **Databricks launched Instructed-Retriever-1.** Instead of sequential agentic search loops, the model scales retrieval in parallel by generating multiple query and filter variants, then reranking them. Databricks said this cuts search time by **more than 3x**, halves answer time, and matches Claude Sonnet 4.5 retrieval quality on KARLBench. [20]
- **GitHub Copilot expanded to a 1M-token window.** Copilot now supports a **1 million** context window and configurable reasoning levels

for VS Code, Copilot CLI, and app developers. [21]

Industry Moves

Why it matters: companies are increasingly selling measurable outcomes, broad AI access, and long-term platform bets—not just model access.

- **Cognition put a financial guarantee behind Devin.** Its new AI Productivity Guarantee says that if Devin delivers less engineering value than customers pay for, Cognition will fund usage until it does, up to **$10 million**. The company also published how it estimates productive output and human-equivalent engineering time. [22, 23, 24]
- **Perplexity partnered with the U.S. Small Business Administration on a mass adoption push.** The Main Street AI Accelerator will provide **$25M** in compute credits—**$250** each for up to **100,000** eligible companies. [25]
- **GeneralistAI raised $400M.** The company said the new capital will go toward building general intelligence for the physical world and making it useful to everyone. [26]

Policy & Regulation

Why it matters: biosecurity and national AI policy both moved closer to concrete action.

- **A broad coalition urged Congress to mandate DNA synthesis screening.** Signatories including Sam Altman, Dario Amodei, Demis Hassabis, Mustafa Suleyman, Nobel laureates, and DNA-synthesis firms called for mandatory screening and recordkeeping for synthetic nucleic acid orders and the machines that print them, arguing AI is eroding historical knowledge barriers around biological weapons. [27, 28]
- **Canada launched a new national AI strategy.** The government framed AI For All around Canadian values, public accountability, and AI that serves all Canadians; related posts described it as part of building, training, and scaling AI domestically. [29, 30]

Quick Takes

Why it matters: a few smaller updates still sharpened the picture.

- OpenAI said one of its models found a counterexample to an **80-year-old Erdős conjecture** and discussed the discovery on the OpenAI Podcast. [31]
- OpenAI added moderation scores to the Responses API and Completions API so developers can log, route, review, or block within the same request flow. [32]
- ParseBench debuted at CVPR 2026 with **2,000+** enterprise document pages and **167K+** test rules for VLM document understanding. [33, 34]

- Runway said token consumption grew **50%**, power users **140%**, and enterprise NDR reached **300%** in the past six weeks. [35]

Sources

1. X post by @AnthropicAI
2. X post by @AnthropicAI
3. X post by @alexalbert___
4. X post by @kimmonismus
5. X post by @kimmonismus
6. X post by @kimmonismus
7. X post by @vllm_project
8. X post by @NVIDIAAI
9. X post by @ArtificialAnlys
10. X post by @arena
11. X post by @dair_ai
12. X post by @skalskip92
13. X post by @skalskip92
14. X post by @skalskip92
15. X post by @santiaranguri
16. X post by @OpenAI
17. X post by @OpenAI
18. X post by @OpenAI
19. X post by @ChristinaHartW
20. X post by @DbrxMosaicAI
21. X post by @pierceboggan
22. X post by @cognition
23. X post by @cognition
24. X post by @cognition
25. X post by @perplexity_ai
26. X post by @GeneralistAI
27. X post by @TheRundownAI
28. X post by @kimmonismus
29. X post by @MarkJCarney
30. X post by @aidangomez
31. X post by @OpenAI
32. X post by @OpenAIDevs
33. X post by @jerryjliu0
34. X post by @llama_index
35. X post by @c_valenzuelab