Composer 2.5, NVFP4 Pretraining, and Anthropic's Stainless Deal

AI High Signal Digest

2026-05-19

Composer 2.5, NVFP4 Pretraining, and Anthropic's Stainless Deal

By AI High Signal Digest • May 19, 2026

Cursor deepened its model push with Composer 2.5 and a larger SpaceXAI training effort, while NVIDIA showed 4-bit pretraining at scale and Anthropic bought Stainless. Also inside: Meta's AIRA, Devin Auto-Triage, and new infrastructure moves from DecartAI, Hugging Face, Dell, and Zyphra.

Top Stories

Why it matters: today's biggest signals were a new push in coding models, a major efficiency step in pretraining, and tighter control of the developer tooling stack.

- **Cursor launched Composer 2.5** as its most powerful model, saying it is better on long-running tasks and complex instructions, built on Moonshot's Kimi K2.5 base, and up to **10x more efficient** than similarly capable models. Cursor also said it is training a significantly larger model from scratch with SpaceXAI using **10x more compute**, showing a leading coding product pairing distribution with direct model training [1, 2, 3, 4]
- **NVIDIA reported the first public multi-trillion-token 4-bit pretraining run**, training a 12B hybrid Mamba-Transformer on **10T tokens** in NVFP4 while keeping downstream accuracy close to FP8 baselines. An NVIDIA researcher added that **Nemotron 3 Super (120B-12A)** and Nemotron 3 Ultra were also pretrained in NVFP4, suggesting ultra-low-precision pretraining is moving into core model development [5, 6]
- **Anthropic is acquiring Stainless**, the SDK and MCP server platform it said has powered every Anthropic SDK since the earliest days of its API. The deal shows how developer tooling is becoming part of the model

platform itself [7, 8]

Research & Innovation

Why it matters: the clearest research trend is that better system design can matter as much as raw model size.

- **Meta's AIRA** autonomously discovered neural architectures that beat Llama 3.2 at 350M, 1B, and 3B scales under a 24-hour compute budget. It splits work between **AIRA-Compose** for macro architecture and **AIRA-Design** for low-level mechanisms, and that split reportedly beat a single end-to-end agent on real search tasks [9]
- **GPT-5.4 nano plus a critic-comparator loop** reached **76.4%** on SWE-bench Verified, matching Gemini 3 Pro and Claude Opus 4.5 Thinking, by selecting among **k=8** weak-model proposals using execution and proof signals. The reported lesson is that the right patch is often already in the candidate set; the bottleneck is selector quality [10]
- A new **Stanford paper** argues that, on multi-hop reasoning tasks, single-agent systems usually match or beat sequential, debate, role-based, and ensemble multi-agent setups when thinking-token budgets are equal. The explanation given is that every message handoff compresses the reasoning chain and can drop information [11]

Products & Launches

Why it matters: new releases are pushing agents beyond chat into persistent, remote, and proactive workflows.

- **Cognition released Devin Auto-Triage publicly.** Devin monitors bugs, alerts, and incidents, connects to systems like Slack, Linear, GitHub, and observability tools, and returns context, next steps, or a PR; it also remembers prior investigations and recurring issues [12, 13, 14]
- **Claude Code Fast mode** now defaults to **Opus 4.7**, with Anthropic saying it delivers identical quality at about **2.5x** the response speed, at a higher per-token price for latency-sensitive work [15, 16]
- **GitHub Copilot CLI and code sessions** now support **remote control** generally available, letting users monitor progress, approve actions, and respond to prompts from anywhere [17, 18]

Industry Moves

Why it matters: capital and partnerships continue flowing toward the infrastructure layer beneath agents and applications.

- **DecartAI raised $300M in Series B**, bringing total funding above $450M, and launched **DOS 2.0**, which it says delivers over **1,600 tokens/sec** for agentic inference and over **100 FPS** for world models. The

company also said new world models, **Lucy** and **Oasis**, are coming in the next few weeks [19]

- **Hugging Face CEO Clement Delangue said Hugging Face and Dell are collaborating** to bring on-prem and local AI based on open-source models to enterprise, arguing this can help address GPU shortages and offer cheaper, faster, and safer alternatives to cloud APIs [20]
- **ZyphraAI** published end-to-end inference benchmarks on **AMD Instinct MI355X**, saying its optimizations outperformed AMD's baseline and narrowed the gap to NVIDIA B200 for serving Kimi K2.6, GLM 5.1, and DeepSeek V3.2 [21, 22, 23]

Quick Takes

Why it matters: a few smaller updates still sharpened the picture on local AI, benchmarks, robotics, and media generation.

- **Qwen3.7 Preview** reached **#13** in Text Arena and **#16** in Vision Arena, making Alibaba the **#6** lab in text and **#5** in vision [24]
- **llama.cpp** added **MTP** for the **Qwen3.6** family; its maintainers described the performance jump as a major boost for local inference on commodity hardware [25, 26]
- **Figure's humanoids** crossed **119 consecutive hours** of fully autonomous operation and **149,000 sorted packages** [27, 28]
- **Baseten** reported sub-second image generation on B200: **0.98s** for Flux.2 [dev] and **0.87s** for Qwen-Image [29]

Sources

1. X post by @cursor_ai
2. X post by @cursor_ai
3. X post by @cursor_ai
4. X post by @cursor_ai
5. X post by @Marktechpost
6. X post by @ctnzr
7. X post by @AnthropicAI
8. X post by @StainlessAPI
9. X post by @omarsar0
10. X post by @dair_ai
11. X post by @rohanpaul_ai
12. X post by @cognition
13. X post by @cognition
14. X post by @cognition
15. X post by @ClaudeDevs
16. X post by @ClaudeDevs
17. X post by @github

18. X post by @code
19. X post by @DecartAI
20. X post by @ClementDelangue
21. X post by @ZyphraAI
22. X post by @ZyphraAI
23. X post by @ZyphraAI
24. X post by @arena
25. X post by @ggerganov
26. X post by @mervenoyann
27. X post by @Figure_robot
28. X post by @adcock_brett
29. X post by @baseten