Anthropic's Lawsuit, Enterprise Agent Moves, and a $1.03B World-Model Bet

AI News Digest

2026-03-10

Anthropic's Lawsuit, Enterprise Agent Moves, and a $1.03B World-Model Bet

By AI News Digest • March 10, 2026

Anthropic's dispute with the U.S. government escalated as OpenAI secured classified defense access, while Microsoft and OpenAI made new agent moves. A well-funded AMI Labs launch and fresh coding-agent research showed both rapid progress and persistent reliability gaps.

The defense AI dispute turned into a legal fight

Anthropic sued after a federal cutoff, while OpenAI gained classified access

The federal government said it would stop working with Anthropic and designate the company a supply chain risk after Anthropic refused to remove safeguards against mass domestic surveillance and fully autonomous weapons [1]. Anthropic has now filed suit against the Trump administration over the designation [2], while OpenAI separately reached an agreement to have its models used in classified Defense Department settings [1].

> "We cannot in good conscience accede to their request." [1]

Why it matters: A debate that had mostly sat in AI-safety policy is now directly shaping procurement, access, and legal strategy [2, 1]. Anthropic's filing also exposed the business stakes: the company says it has generated more than $5B in commercial revenue, spent $10B on training and inference, and already saw one $15M deal pause after the designation [3].

Agents are moving deeper into enterprise workflows — and into their control stacks

Microsoft launched Copilot Cowork for Microsoft 365

Microsoft introduced Copilot Cowork as a new way to hand off tasks inside Microsoft 365: it turns a request into a plan and executes it across apps and files, grounded in work data and operating within M365 security and governance boundaries [4].

Why it matters: This is a clear signal that agentic task execution is moving into the core productivity suite many enterprises already use [4].

OpenAI is buying Promptfoo to strengthen agent evaluation

OpenAI said it is acquiring Promptfoo, and that Promptfoo's technology will strengthen agentic security testing and evaluation capabilities in OpenAI Frontier [5]. Promptfoo will remain open source under its current license, and OpenAI says it will continue servicing and supporting current customers [5].

Why it matters: As agents get pushed into more real workflows, labs are treating evaluation and security tooling as strategic infrastructure [5].

Research showed both acceleration and friction in AI-for-AI

ByteDance's CUDA Agent pushed low-level automation forward

Researchers from ByteDance and Tsinghua described CUDA Agent, a fine-tuned Seed 1.6 model built for GPU programming, trained on a 6,000-sample operator dataset and run in an agent loop with tools for profiling, editing, compiling, and evaluation [6]. They report that it beats `torch.compile` on 100% of Level-1 and Level-2 KernelBench tasks and 92% of Level-3 tasks, roughly 40% ahead of Claude Opus 4.5 and Gemini 3 Pro on Level-3 [6].

Why it matters: This is a concrete example of AI improving the software stack beneath AI itself. It arrives alongside new work from GovAI and Oxford proposing 14 metrics for tracking AI R&D automation and oversight [6], and Ajeya Cotra's view that software-agent time horizons are moving faster than she expected earlier this year [6].

But long-horizon maintenance and reproducibility are still weak

The split-screen was sharp. SWE-CI tracks code maintenance over 71 consecutive commits, and testing across 100 real codebases over 233 days reportedly found that 75% of models broke previously working code during maintenance; only Claude Opus 4.5 and 4.6 stayed above a 50% zero-regression rate [7]. Separately, an arXiv preprint auditing shadow APIs that claimed GPT-5 or Gemini

access found 187 papers using them, with performance divergence up to 47% and 45% fingerprint-test failures [8].

Why it matters: Strong results on narrow optimization tasks do not remove harder problems around sustained maintenance, trustworthy model identity, and reproducible research [7, 8].

A large new bet formed around world models and physical AI

AMI Labs launched with $1.03B and a world-model agenda

AMI Labs launched with Saining Xie and Yann LeCun, saying it is building AI systems centered on world models that understand the world, retain persistent memory, reason and plan, and remain controllable and safe [9, 10]. The company said it raised $1.03B and is operating from Paris, New York, Montreal, and Singapore from day one [10].

Why it matters: This is a large capital commitment behind an alternative frontier agenda that emphasizes world understanding, memory, planning, and control [10].

ABB and NVIDIA turned physical AI into a more concrete factory software story

ABB Robotics and NVIDIA said they are integrating Omniverse libraries into RobotStudio to launch RobotStudio HyperReality in the second half of 2026 [11]. The companies say the system can reach 99% sim-to-real correlation, cut deployment costs by up to 40%, accelerate time to market by up to 50%, and reduce setup and commissioning times by up to 80%, with Foxconn and Workr already piloting it [11].

Why it matters: Physical AI is becoming a real industrial software stack, not just a research theme [11]. The framing lines up with Fei-Fei Li's argument that "spatial intelligence" — linking perception, reasoning, and action in 3D and 4D worlds — is the next frontier [12].

Sources

1. Anthropic and Alignment | Stratechery by Ben Thompson
2. X post by @Hadas_Gold
3. X post by @ZeffMax
4. X post by @satyanadella
5. X post by @OpenAI
6. Import AI 448: AI R&D; Bytedance's CUDA-writing agent; on-device satellite AI
7. X post by @ChrisLaubAI

8. r/MachineLearning post by u/Electrical-Shape-266
9. X post by @sainingxie
10. X post by @amilabs
11. ABB Robotics Taps NVIDIA Omniverse to Deliver Industrial-Grade Physical AI at Scale
12. Fei-Fei Li on the Future of AI | Human-Centred AI & Spatial Intelligence | India AI Impact Summit