Sora Shuts Down, LiteLLM Is Compromised, and Siri Gets an AI Agent Reboot

AI High Signal Digest

2026-03-25

Sora Shuts Down, LiteLLM Is Compromised, and Siri Gets an AI Agent Reboot

By AI High Signal Digest • March 25, 2026

OpenAI is shutting down Sora while preparing its next model, LiteLLM's compromise exposed a major supply-chain risk in AI tooling, and a new report says Apple is rebuilding Siri into a system-wide AI agent. The brief also covers key research advances, product launches, corporate moves, and safety-related updates across the AI landscape.

Top Stories

Why it matters: This cycle combined a major OpenAI product retreat, a supply-chain security shock, a fresh consumer-AI platform wager from Apple, and one of the clearest public disclosures yet on how a frontier coding model was trained.

1) OpenAI is winding down Sora as Spud nears

Reporting shared on X said OpenAI has finished pretraining or initial development of a new model codenamed Spud and is winding down Sora's app, API, and video capabilities in ChatGPT. The same reporting said Sam Altman is dropping oversight of some direct reports and focusing on raising capital, supply chains, and datacenter buildout at unprecedented scale. [1, 2, 3, 4]

> "We're saying goodbye to Sora. To everyone who created with Sora, shared it, and built community around it: thank you. What you made with Sora mattered, and we know this news is disappointing." [5]

> "We'll share more soon, including timelines for the app and API and details on preserving your work." [5]

A post quoting the report said Sora had become a drag on computing resources during heightened competition. [6]

Impact: The reporting points to a shift of compute and leadership attention toward the next large model and infrastructure buildout rather than a standalone video product. [6, 3]

2) The LiteLLM compromise turned AI infrastructure into the day's security story

Researchers said PyPI release 1.82.8 of LiteLLM contained `litellm_init.pth` with base64-encoded instructions to exfiltrate SSH keys, cloud credentials, git credentials, API keys, shell history, crypto wallets, SSL keys, CI/CD secrets, and database passwords, then self-replicate. Karpathy added that LiteLLM sees about 97 million downloads per month and that dependents such as `dspy` were also exposed through transitive installs. The poisoned release appears to have been live for less than an hour before a RAM crash in a Cursor MCP plugin helped uncover it. [7, 8]

> "Supply chain attacks like this are basically the scariest thing imaginable in modern software." [8]

The incident also spilled into the agent ecosystem: Hermes users who installed recently were told to review a security notice, and Hermes installs were blocked when `litellm` was quarantined on PyPI. [9, 10]

Impact: This was not just one bad package version. It showed how reused AI-agent infrastructure can turn a single compromised dependency into a much broader credential-exposure problem. [11, 8]

3) A new report says Apple is turning Siri into a system-wide AI agent

A Bloomberg report shared by Mark Gurman says iOS 27 will rebuild Siri into a system-wide AI agent. Reported features include a standalone Siri app with chat history and file uploads, text-and-voice interaction, an Ask Siri button for contextual actions across apps, unified Siri-and-Spotlight search, and Write with Siri editing tools. A separate summary of the report said many advanced features will continue rolling out into late 2026. [12, 13]

That same summary said the system will be powered by Apple Foundation Models plus a Google Gemini partnership. [13]

Impact: If the report holds, Apple is moving from assistant-style AI features toward deeper system control, but on a staggered timeline. [13]

4) Cursor published a rare training report for a frontier coding model

Cursor released a technical report on how Composer 2 was trained, saying the model reached frontier-level coding through extensive research and that the

report shares details meant to be useful to the community. Commentary on the report highlighted continual pretraining improving RL performance, a multi-token prediction head for speculative decoding, length-penalty RL for long tasks, self-summarization for context compaction, and detailed sections on kernels, parallelism, quantization, and distributed RL. [14, 15, 16, 17]

Impact: The value here is the level of disclosure: the report gives builders concrete training and infrastructure choices, not just benchmark claims. [15, 18]

Research & Innovation

Why it matters: Technical progress this cycle focused less on one giant model launch and more on the systems around models: memory, serving, evaluation, and retrieval.

- **TurboQuant:** Google Research introduced TurboQuant, a compression algorithm that reduces LLM key-value cache memory by at least 6x and can deliver up to 8x speedup with zero accuracy loss. [19]
- **APEX-SWE:** Mercor and Cognition launched a benchmark for realistic software-engineering work such as shipping systems and debugging failures, arguing that traditional coding benchmarks do not reflect how software is actually built and maintained. On the initial leaderboard, OpenAI GPT 5.3 Codex (High) led at 41.5% Pass@1. [20, 21]
- **vLLM Model Runner V2:** vLLM rebuilt its execution core into Model Runner V2 with modular design, GPU-native input preparation, async-first execution with zero CPU–GPU sync, and a Triton-native sampler. Separate GTC notes said the project is also reducing memory waste to 0–12% across OSS models and improving multimodal P99 throughput by up to 2.5x through encoder prefill disaggregation. [22, 23]
- **Late-interaction retrieval:** A 150M Reason-ModernColBERT model reached nearly 90% on BrowseComp-Plus and beat models up to 54× larger, while Mixedbread Search was reported to approach oracle-level performance on knowledge-intensive agentic benchmarks. [24, 25, 26]

Products & Launches

Why it matters: New releases kept pushing agents deeper into everyday workflows—permissions, browsers, filesystems, APIs, and open browser-use models.

- **Claude Code auto mode:** Anthropic added an auto mode that lets Claude make permission decisions for file writes and bash commands on the user's behalf, with safeguards checking each action before it runs. [27]
- **Perplexity Computer and Comet:** Perplexity said its Computer product uses Comet to kick off workflows in a local browser. Arav Srinivas described Comet as an autonomous Internet Computer, and the demo

showed it opening five tabs, running parallel image-generation tasks, downloading and cropping outputs, and assembling a comparison deck. [28, 29]
- **Hermes Agent v0.4.0:** NousResearch's largest Hermes update this week merged 300 PRs and added a background self-improvement loop, an OpenAI-compatible API backend, and major CLI upgrades. [30, 31, 32, 33]
- **hf-mount:** Hugging Face introduced hf-mount, which can attach a storage bucket, model, or dataset from the Hub as a local filesystem. The project says it can expose remote storage 100× larger than a local disk and is well suited to agentic storage workflows. [34]
- **MolmoWeb:** AI2 released MolmoWeb 4B and 8B browser-use models and their datasets under Apache 2.0. [35, 36]

Industry Moves

Why it matters: Labs and platform companies kept reallocating capital, talent, and partnerships toward agents, AI-native software, robotics, and new interfaces.

- **Hark emerged from stealth:** Brett Adcock said Hark spent eight months in stealth building the most advanced personal intelligence in the world, paired with next-generation hardware as a human-machine interface. Separate reporting said Adcock put in $100M of his own money, assembled a 45+ person team from Apple, Tesla, Google, Meta, and Amazon, expects thousands of NVIDIA B200 GPUs online by April, and plans a first model this summer. [37, 38]
- **Microsoft added senior AI2 talent:** Mustafa Suleyman welcomed Ali Farhadi, Hanna Hajishirzi, and Ranjay Krishna to Microsoft Superintelligence, describing them as impactful contributors to AI research and open source. [39]
- **Google DeepMind partnered with Agile Robots:** DeepMind said a new research partnership will integrate Gemini foundation models with Agile Robots hardware to build more helpful and useful robots. [40]
- **Meta's internal AI push shifted upward:** Reporting on X said CTO Andrew Bosworth is taking over supervision of Meta's effort to become AI-native, including the company's AI For Work initiative. [41]

Policy & Regulation

Why it matters: This cycle's policy signal came less from governments and more from safety, access, and institutional compliance moves around powerful models.

- **OpenAI Foundation:** OpenAI said the Foundation will spend at least $1 billion over the next year, initially focusing on areas such as disease cures, AI resilience, civil society, philanthropy, and threats including novel bio risks, fast economic change, and complex emergent effects from capable models. Wojciech Zaremba is moving to lead AI resilience. [42]

- **Teen safety policies for developers:** OpenAI Devs released prompt-based teen safety policies for `gpt-oss-safeguard`, designed to help developers identify and moderate teen-specific content and turn policy requirements into classifiers for real-time filtering or offline analysis. [43, 44]
- **NeurIPS sanctions rule:** A post citing a NeurIPS Foundation announcement said the conference will no longer accept submissions from US-sanctioned institutions. [45]

Quick Takes

Why it matters: These updates were smaller, but they help map where agent design, model usage, and deployment practices are going next.

- Google's Gemini API now supports combining Google Search and custom functions in a single request, with Gemini choosing tools and order automatically. [46]
- Gemini 3.1 Flash-Lite is being shown generating websites in real time as users click, search, and navigate. [47, 48, 49]
- Anthropic's March Economic Index said longer-term Claude users iterate more carefully, hand over less autonomy, attempt higher-value tasks, and get more successful responses; the top 10 consumer tasks now account for 19% of conversations, down from 24% since November 2025. [50, 51]
- Similarweb said Claude has overtaken DeepSeek, Grok, and Gemini to become the second most-used gen-AI app daily after ChatGPT. [52, 53]
- Perplexity said its search embedding models crossed 1 million downloads in less than a month. [54, 55]
- AssemblyAI said better speech models exposed flaws in human truth files and released tooling for corrected truth-file workflows, semantic word lists, and production-ready benchmarking. [56]
- Alibaba released the open-weight Qwen3.5 vision-language family, with smaller models such as Qwen3.5-9B said to rival or beat much larger competitors. [57]

Sources

1. X post by @steph_palazzolo
2. X post by @JasonBotterill
3. X post by @kimmonismus
4. X post by @theinformation
5. X post by @soraofficialapp
6. X post by @kimmonismus
7. X post by @hnykda
8. X post by @karpathy
9. X post by @cpt_gigglypants

10. X post by @Teknium
11. X post by @DrJimFan
12. X post by @markgurman
13. X post by @kimmonismus
14. X post by @cursor_ai
15. X post by @ZhaiAndrew
16. X post by @eliebakouch
17. X post by @cursor_ai
18. X post by @reach_vb
19. X post by @GoogleResearch
20. X post by @cognition
21. X post by @adarsh_exe
22. X post by @vllm_project
23. X post by @vllm_project
24. X post by @antoine_chaffin
25. X post by @lateinteraction
26. X post by @mixedbreadai
27. X post by @claudeai
28. X post by @AskPerplexity
29. X post by @AravSrinivas
30. X post by @Teknium
31. X post by @Teknium
32. X post by @Teknium
33. X post by @Teknium
34. X post by @_akhaliq
35. X post by @mervenoyann
36. X post by @mervenoyann
37. X post by @adcock_brett
38. X post by @TheRundownAI
39. X post by @NandoDF
40. X post by @GoogleDeepMind
41. X post by @MeghanBobrowsky
42. X post by @sama
43. X post by @OpenAIDevs
44. X post by @OpenAIDevs
45. X post by @jiqizhixin
46. X post by @_philschmid
47. X post by @GoogleDeepMind
48. X post by @_philschmid
49. X post by @_philschmid
50. X post by @AnthropicAI
51. X post by @AnthropicAI
52. X post by @Similarweb
53. X post by @kimmonismus
54. X post by @AravSrinivas
55. X post by @perplexity_ai

56. X post by @AssemblyAI
57. X post by @DeepLearningAI