

White House Forms a "Super Intelligence Force" as OpenAI's Capacity Squeeze Pushes Users Toward Claude

AI High Signal Digest

2026-10-05

White House Forms a "Super Intelligence Force" as OpenAI's Capacity Squeeze Pushes Users To- ward Claude

By AI High Signal Digest • October 5, 2026

A new White House body to coordinate federal superintelligence work, a reported US open-weight push led by Reflection AI, rising frustration over OpenAI's usage cuts, and research on verifying and scaling agents.

Washington creates a superintelligence coordination body

The White House posted a statement from President Trump announcing a "Super Intelligence Force (SIF)". Its job is to coordinate the federal government's effort "to ensure that America continues to lead the World in Super Intelligence" [1]. The post gives no structure, budget or leadership.

A second story points in the same direction. Axios reports, as relayed on X, that Reflection AI is about to release an "extremely capable" open-weight model expected to compete with the top Chinese open models. The relaying post says Reflection has paid \$150M a month for compute at Colossus since July, on top of a \$1B compute deal with Nebius [2]. Axios's sources also say several other unnamed American labs plan open releases this month. The US government wants American open models to compete globally, and Howard Lutnick reportedly once considered funding them directly [2]. @teortaxesTex is skeptical: Nvidia and Thinky have so far failed to match even GLM 5.3. He adds that a lull in frontier Chinese releases gives Reflection an opening [3]. On the Chinese side, Alibaba has confirmed Qwen 4 is in training but has given no date. Kimi K3.1 and next-GLM timing is still speculation [4].

OpenAI’s compute squeeze has users talking about switching

Prominent users report that OpenAI suspended new sign-ups for its \$200 plan. They say it then effectively halved usage limits across all plans and offered GPT-6.1 Sol as the more efficient alternative. One critic puts this down to severe compute constraints and faults OpenAI for presenting the cut as an efficiency gain [5]. Theo argues the coding gap has widened since July. In his view, Opus 5.5 is now fast, cheap and nearly limitless on a \$200 plan, while OpenAI models are “slow as hell without fast mode” and can burn through a \$200 plan in hours [6]. OpenAI’s Thibault Sottiaux replied with a commitment: for the next 28 days, Codex will ship either a clear improvement or a full limit reset every day [7]. Critics read this as an attempt to stop users defecting to Claude after a thin DevDay [8].

Not all the evidence cuts against OpenAI. One user says Codex limits feel fine with 6.1 Sol as a daily driver. He cites Arena cost-per-task data showing about $5\times$ more usage than Astra [9]. GPT-6 Astra also took #1 on four Design Arena leaderboards, including 3D Design (1484) and Frontend (1397), about a month after release [10]. Supply is clearly the constraint. One unverified post says OpenAI’s ultrafast tier runs on Cerebras chips, and that supply ran short because Jane Street bought it up, reportedly at \$200M per megawatt [11].

Deals and infrastructure

- **AMD–World Labs:** AMD is buying Fei-Fei Li’s World Labs for \$8.2B and naming her EVP and chief scientist. That gives AMD a world-model answer to Nvidia’s Cosmos [12].
- **Nvidia neutrality:** SemiAnalysis says Nvidia’s support for non-Nvidia chips in SLURM has worsened since it acquired SchedMD, despite its neutrality pledge. AMD built its own competing scheduler, “spur.” SemiAnalysis asks whether Hugging Face, also acquired by Nvidia, will go the same way [13].
- **Agent guardrails in hardware:** Nvidia’s Open Agent Safety Platform combines OpenShell access control with Sentry monitoring on BlueField-4 DPUs. This moves guardrails off the processor the agent runs on [14].
- **China:** Alibaba T-Head unveiled the Zhenwu V900, with 216 GB of memory and 1,200 GB/s of interconnect bandwidth. It is billed as “3x” the M890 and ships from Q1 2027 [15]. Filings reviewed by Bloomberg show a Chinese financing firm controlled by local government entities funded purchases of more than 700 servers. One contract covers 32 Asus servers with B300 chips, which need a US license for sale to China [16]. @teortaxesTex calls such leaks “microscopic.” He argues that what counts is domestic production and offshore contracted capacity, mainly in Southeast Asia [17].

Agent research: verification, compression and many agents at once

This period’s papers keep returning to one question: where should extra compute go around a fixed model? - **NVIDIA Mid-Harness:** sample several candidate shell commands and verify them before running one. With a GPT-5.6 Sol verifier choosing among 8 actions, TerminalBench-Lite Pass@1 rose from 50% to 68%. With a weak verifier, extra samples added almost nothing [18]. - **Google VeriHarness:** agreement among rollouts can hide shared errors. The same base model acts as a verifier that settles disagreements against workspace evidence and challenges answers every rollout agreed on. This added 6.2 points over a single rollout with Gemini 3.5 Flash and 6.4 with Claude Opus 4.8. The authors also released about 26,000 rollouts [19]. - **UT Austin compression study:** across about 35,000 runs, compression policies that used about a third of the tokens took 20–80% longer on Terminal-Bench with Qwen. A policy tuned for Qwen dropped Devstral to 38.7% [20]. - **Meta CLMs:** the model edits its own live context. With no training, this gave 11.4% higher accuracy using 21.5% fewer FLOPs on BrowseComp-Plus. Suffix Cache Reuse cuts server compute by 35% [21]. - **Microsoft Agensh:** a multi-agent harness with no central orchestrator, run with up to 1,024 coding agents. Going from 1 to 1,024 agents raised the test-pass rate for building pandoc from scratch from 33.89% to 55.06% [21]. - **EverMind Raven** (open source): evolves a separate harness for each model and domain. On DeepSeek-V4-Flash it scored 69.3% on BrowseComp and 60.0% on Humanity’s Last Exam, against at most 62.4% and 43.3% for two other harnesses on the same model [22].

Generative media and agent-driven reverse engineering

Kling 4.0 Flash is in early access, with the full launch due in October [23]. Kling says it supports native 30-second generation, up to 15 multimodal references and 10 keyframes [23]. A widely shared essay lists projects where agents decompiled games. Claude Code agents matched all 1,446 functions of *LSD: Dream Emulator* in 32 days. Up to 17 Claude agents have matched about 70% of *Modern Warfare 2*’s functions in two months [24]. The essay notes that distributing the results is legally contested: Take-Two sued over a GTA III decompilation, and Activision objected to the MW2 project [24].

Safety and the consciousness argument

Following last period’s resignation of OpenAI’s David Robinson, former OpenAI researcher Ryan Lowe said attempts since about 2021 to establish “systems safety” at OpenAI never took root. He called for layered safety drawn from fields like nuclear energy, which he says needs a shift in incentives at labs including Anthropic [25]. The consciousness debate continued too. Mustafa Suleyman argued that Claude’s uncertainty about being conscious reflects Anthropic’s training choices rather than evidence, and could mislead users [26]. Elsewhere,

Cline paused its free DeepSeek-V4.1-Flash promotion, citing abnormally high abuse [27].

Sources

1. X post by @WhiteHouse
2. X post by @AndrewCurran_
3. X post by @teortaxesTex
4. X post by @kimmonismus
5. X post by @kimmonismus
6. X post by @theo
7. X post by @thsottiaux
8. X post by @kimmonismus
9. X post by @petergostev
10. X post by @DesignArena
11. X post by @cryptopunk7213
12. X post by @dl_weekly
13. X post by @SemiAnalysis_
14. X post by @dl_weekly
15. X post by @SemiAnalysis_
16. X post by @poezhao0605
17. X post by @teortaxesTex
18. X post by @dair_ai
19. X post by @omarsar0
20. X post by @omarsar0
21. X article by @dair_ai
22. X post by @dair_ai
23. X article by @Kling_ai
24. X article by @rehan_shei
25. X post by @ryan_t_lowe
26. X post by @kimmonismus
27. X post by @cline