Jensen Huang Resigns as Nvidia CEO While Forensic Claims of Blackwell-Trained Chinese Models Escalate the Export-Control Fight

AI High Signal Digest

2026-11-13

Jensen Huang Resigns as Nvidia CEO While Forensic Claims of Blackwell-Trained Chinese Models Escalate the Export-Control Fight

By AI High Signal Digest • November 13, 2026

Nvidia's leadership upheaval and a contested weight-forensics report are driving a new push for US chip export controls. Google ships Gemini 4 Hydrogen, and Thinking Machines open-sources Inkling 2 Small.

Huang is out at Nvidia, and export-control politics move fast

Jensen Huang has resigned as Nvidia's President, CEO and board member, effective immediately [1]. CFO Colette Kress takes over as interim CEO [2]. Bloomberg, as relayed in a post, reports that the board met late into the night and left him no choice but to resign [3]. The NYT reports he has hired the law firm Paul Weiss [4]. According to WIRED, an internal post asking the board to weigh employee sentiment drew more than 1,600 comments, most of them backing his return [5]. CNN sources say US intelligence on the Chinese businesswoman linked to Huang does not so far prove she was spying for Beijing [6].

The policy effects matter more than the personnel change. Shakeel Hashim points out that Nvidia was leading the lobbying against the NDAA export-control measures, and argues that its disarray and lost credibility mean it can no longer do so [7]. Andrew Curran reports that OpenAI is joining Anthropic in backing bipartisan chip export controls, with the aim of passing legislation before year-end [8]. Separately, The Information reports that Litong Inc, a group of companies tied to Alice Huang, has imported 240,000 Nvidia chips

worth at least $7.5B into China since 2023 despite export controls. It says the group also invested in DeepSeek, which is still trying to get more Blackwell chips [9].

Commerce is also reportedly weighing a rule on China's access to AI chips in overseas data centers. One analysis says the rule as described has three gaps:

- it covers only certain countries;
- it applies only to future exports, so existing data centers are untouched;
- it only covers chips at least as powerful as Blackwell, so H200-class chips stay available.

The analysis puts the effect at no more than 14% of China's end-2027 compute, and calls that "almost certainly a strong ceiling" [10].

Contested forensics: were Chinese frontier models built on Blackwell?

Will Depue says checkpoint forensics on released weights shows that DeepSeek, Moonshot, MiniMax and Ant Group built their newest models on Blackwell chips, which are barred from export to China [11]. His method relies on a claim that each GPU-and-seed combination leaves an initialization pattern that survives training [11]. He searched more than six trillion combinations and says six Chinese models, including DeepSeek-V4-Pro, Kimi K3 and MiniMax-M3, match Blackwell. Four of them match GB200/GB300 parts specifically [11].

His evidence and estimates:

- **Reproduction tests:** a B300 reproduced DeepSeek-V4-Flash's starting weights bit for bit, and a GB200 reproduced Kimi K3's to 0.999995 correlation. An H100 did not reproduce the DeepSeek weights [11].
- **Cluster size:** he estimates Kimi K3 and DeepSeek-V4-Pro each used 2,000–3,000 Blackwell GPUs, at roughly $150–200M per cluster [11].
- **Outside checks:** he says senior OpenAI and Anthropic researchers independently verified the results [11].
- **Location:** he concedes the method cannot show where the chips physically are [12].

The pushback came quickly:

- **Initialization isn't training.** Dfranke argues that initialization is computationally trivial and could be done on separate hardware, so it says nothing about where a model was trained [13]. Depue answers that the fingerprint matched publicly disclosed hardware in all 24 known cases. He adds that Kimi K3's weights were created by at least 64 processes, all on Grace Blackwell [14].
- **No match for DeepSeek's latest model.** Depue acknowledges that DeepSeek-V4.1-Flash produced no match. He says this means DeepSeek changed how it creates starting weights, which reveals nothing about the

chip either way [15]. Late in the period he posted that it "might be NVIDIA," without further detail [16].
- **The scale is small.** Teortaxes notes that against roughly 24 GW of installed Chinese compute, even five labs with 10 MW each come to 0.2%. He calls that a success for export controls, not a failure [17].

Depue has released his code [18]. Treat the hardware attribution as plausible but unconfirmed. It is already shaping the policy debate.

Gemini 4 Hydrogen: Google's fast tier, with independent caveats

Google launched Gemini 4 Hydrogen as the execution model alongside Carbon and Argon. It is available across AI Studio, the API, the Gemini app, Search and Antigravity [19]. Pricing is a flat \$1.50/\$7.50 per 1M tokens across a 1M-token context window. Google says it scores 51 on the Artificial Analysis Intelligence Index at high reasoning, 54 at xhigh, against Argon's 50 [19]. In Antigravity, a new Orchestrator switches between Carbon and Hydrogen within a session. Google reports near-Carbon success on Terminal-Bench 4.0 while using 24% less quota [19].

Independent evaluations:

- **Artificial Analysis:** puts it 1 point behind Claude Sonnet 5.5 and GPT-6.1 Sol at about 32% of Sonnet's cost but about 2.5× Sol's [20]. It has an 8% hallucination rate [20] and is weaker on terminal-based agentic work (54% on Terminal-Bench 4.0) [20].
- **Vals:** #6 on its index, 2.0 points behind Carbon at 38% of Carbon's list price [21].
- **Agent Arena:** #5, and the fastest model in its top 10 at a median 5.3 minutes per task. Arena notes it spends 41% more output tokens per task than Carbon [22].

Developers raised complaints too:

- **Pricing:** Simon Willison finds the pricing confusing. Hydrogen is only a 25% discount on Carbon and Argon, though those two double in price on Dec 1 [23].
- **Budget exhaustion:** at xhigh, one tester saw Hydrogen spend 131K thinking tokens on reading a jersey number and return no answer. The same budget exhaustion hit 1.8% of their text benchmark [24].

Inkling 2 Small goes open-weight

Thinking Machines released Inkling 2 Small under Apache 2.0 [25]. It is a mixture-of-experts model with 594B total and 17B active parameters, takes text and images as input, and supports up to 1M tokens of context. vLLM says it fits on a single GB200/GB300 node [26]. Arena ranks it the #1 US open

model in Agent Arena and #2 among all open models. Its median cost is $0.26 per task, about 4.7× less than Kimi K3.1 Pro [27].

Other notable items

- **OpenAI cyber:** OpenAI launched two security products, Daybreak Defense Factory and Daybreak Red Team [28]. A slide from the livestream reportedly shows an unreleased OpenAI model at 77% on an unnamed cyber metric, against about 30–50% for Opus 5.5, Mythos 5.2 and Carbon [29].
- **Anthropic:** Self-hosted Claude Managed Agents is now in private beta [30]. Claude Meetings, which builds documents and other artifacts during calls, is rolling out on Team and Enterprise plans [31]. Theo reports the IPO is set for next Wednesday [32].
- **Perplexity Decider v1.3:** Perplexity open-sourced the latest versions of its decision models and added a "locate" span-extraction feature [33]. Internally, Decider read 1B tokens across 7,000 agent sessions in one day, which led to 14 fixes to Perplexity's agent security rules [34].
- **Rubin inference:** vLLM reports DeepSeek-V4.1-Flash on Vera Rubin NVL72 at 2.52M tokens/s/GPU peak, 2.7× GB300 [35].
- **OWE-Bench:** Mercor's new benchmark of 26 long-horizon engineering projects is led by Fable 5.2 at 54.6%, ahead of GPT-6 Astra at 44.2% [36].
- **Speech:** Salesforce's SOV-Transcribe-1 is now the top open-weights speech-to-text model on Artificial Analysis, with 2.3% WER [37].

Sources

1. X post by @kimmonismus
2. X post by @kimmonismus
3. X post by @wallstengine
4. X post by @wallstengine
5. X post by @WIRED
6. X post by @NatashaBertrand
7. X post by @ShakeelHashim
8. X post by @AndrewCurran_
9. X post by @theinformation
10. X post by @fiiiiiist
11. X post by @willdepue
12. X post by @willdepue
13. X post by @dfranke
14. X post by @willdepue
15. X post by @willdepue
16. X post by @willdepue
17. X post by @teortaxesTex
18. X post by @willdepue

19. X article by @GoogleAIStudio
20. X post by @ArtificialAnlys
21. X post by @ValsAI
22. X post by @arena
23. X post by @simonw
24. X post by @skalskip92
25. X post by @thinkymachines
26. X post by @vllm_project
27. X post by @arena
28. X post by @gdb
29. X post by @scaling01
30. X post by @ClaudeDevs
31. X post by @claudeai
32. X post by @theo
33. X post by @AravSrinivas
34. X post by @alexatallah
35. X post by @vllm_project
36. X post by @mercor
37. X post by @ArtificialAnlys