Huang returns to Nvidia as Midha says stranded power and RL compute will reshape AI infrastructure

VC Tech Radar

2026-11-24

Huang returns to Nvidia as Midha says stranded power and RL compute will reshape AI infrastructure

By VC Tech Radar • November 24, 2026

Nvidia's board brought Jensen Huang back after the DOJ said it had no evidence against him. Anjney Midha laid out a compute-and-power thesis that challenges how VCs underwrite. Embo closed a $110M seed for world-model robotics, and agent commerce and model-adoption data kept coming.

Nvidia brings Huang back

Nvidia's board has reinstated Jensen Huang as president, CEO and board director [1]. A WSJ excerpt shared on Reddit explains how it happened. Over the weekend, Attorney General Todd Blanche told Trump and Vance that Huang was not a target or subject of any DOJ investigation, and that the department had no evidence he broke laws or posed a national-security risk. That assessment rested on Nvidia's months of cooperation with DOJ probes into two customers suspected of smuggling AI chips into China [2].

Allie K. Miller puts the stock's low at $125 last Friday, against $213 before news of the affair broke on October 29. She expects close scrutiny of everything Huang says about China for weeks [3]. Amjad Masad called Huang "one of the most important champions of American open-weight AI" [4]. The prior brief carried Jason Lemkin's warning that founders should redo exit math that assumed Nvidia as a buyer. With Huang back, that advice is worth revisiting.

Midha: power, RL compute and "fire the bankers"

On 20VC, Anjney Midha, an early Anthropic investor and founder of AMP, made the period's most detailed infrastructure argument:

- **Stranded power is the opportunity.** He says the US has more than 100 GW of stranded power. Reinforcement learning (RL) and post-training don't need gigawatt-scale clusters, and about 50% of his portfolio's compute spend is now in post-training [5]. AMP co-founded a company building prefabricated 5–10 MW "grid pods" that plug into powered sites [5].
- **Pricing.** He says AMD MI355 capacity on AMP's National Compute Grid clears at $12/hour, against $3–4 on a five-year lease. His argument is that pay-for-completed-work scheduling still lowers total cost [5].
- **Custom silicon.** The grid plans 2 GW of partner-funded capacity, anchored by Zurich, plus 5 GW of third-party capacity by 2030, with about 25% targeted at workload-specific silicon. He says orders for RL-oriented custom chips have started [5].
- **Where he won't invest.** He calls inference "a feature not a standalone service" that serious labs will bring in-house, and is not an investor in that category [5]. He says neoclouds valued like technology companies are "seeing a reckoning," with security his biggest concern [5]. He is bullish on Arena's evaluation business, which he estimates has passed $200M in annualized consumption [5].
- **Underwriting.** He supports tranches tied to execution milestones. He cites OpenRouter's sale to Stripe, reportedly north of $7B, as a more-than-$1B return on his roughly 20% a16z stake within 8 months. He calls tranches without milestones "smoke and mirrors" [5]. He also says 2021 growth marks on payroll SaaS, fintech and crypto "are not real" [6], and that a US cluster of 24,000 was pre-sold to a major Chinese company because neoclouds lacked credit support [6].

These are claims from a single investor with positions in the companies he names, so treat them as a thesis rather than market data.

Early-stage rounds and growth numbers

- **Embo ($110M seed, led by a16z, with Khosla, DST, NVentures and Striker).** Embo pretrains robot "brains" on internet video, then practices motor skills in imagined scenarios [7]. Founders Danijar Hafner and Wilson Yan built Dreamer and VideoGPT [8]. Nathan Benaich backed Hafner after he left DeepMind [9].
- **Noeri ($110M).** A "neolab" from Stanford's Yejin Choi and Carlos Guestrin built around many specialized models trained on organizations' own data. Import AI notes it is light on technical details [10].
- **River.** Its API passed $100M ARR five months after launch and is on track for more than $200M this year. Perplexity was a design partner [11].

- **Exa.** a16z says Exa grew 10x for three straight years, handles 2B requests a month, and raised at a $2.2B valuation this year [12].
- **Sierra.** Bret Taylor's agent company is being valued at over $30B in a new announcement [13].
- **Replit acquires Resonate,** which builds durable infrastructure for agents. Masad says it is the first company he has seen ship with a verifiable spec in Lean and TLA+ [14, 15].

Gemini 4 Neon: adoption data and doubts about the hype

Aravind Srinivas says Neon usage on Perplexity compounded at 10.6% a day in Ask and 15.0% a day in Computer from Nov 19–22 [16]. A later update put Neon up 45.4% day over day in Computer [17]. He says the demand is organic, since Perplexity doesn't yet route traffic to Neon by default, and that it comes mostly from users switching away from Opus 5.5 and Sol 6.2 [18].

Martin Casado likes the model but flagged near-identical praise posts. Several of the accounts list themselves on unfair.so, a marketplace for sponsored X posts. He says this "doesn't prove anything" [19]. In practice, weight Perplexity's usage data well above social chatter.

OpenAI: biology tools, chip design and standards

On a16z, OpenAI said it acquired Orchestra Bio to build a "Codex for scientists." Its reasoning is that models like GPT-Rosalind score well on benchmarks, but scientists don't know how to use them [20]. In one collaboration, 2,000 agents analyzed 8 million cells and found a cell-type-specific gene effect that months of manual analysis had missed. Agents at that scale are not yet generally available [20]. For founders, OpenAI wants specialist tools to plug in through MCP and connectors [20]. It also expects demand for fresh biological data to keep growing, "possibly linearly" [20].

Greg Brockman said OpenAI's models, working with AMD, delivered "a year's worth of full generational improvements in 5 months" in chip design [21]. He added that chain-of-thought monitoring is degrading as models do more work in the forward pass [21], and that compute will be "the biggest limiter" [21].

At the UN, Sam Altman asked for national and international frontier-AI standards that "should not lock in incumbents" [22]. He also claimed an OpenAI model solved the Navier–Stokes Millennium Prize problem [22]. That claim has not been independently confirmed in what we reviewed. OpenAI and Cisco separately open-sourced Shirudo, a runtime that lets agents deploy eBPF shields against exploits without downtime [23].

Agent commerce: narrow shortlists, guest checkout wins

a16z and The Prompting Company ran 11,307 agent tests across 1,270 prompts [24]. Main findings:

- **Agents converge on the same products.** A product picked by one agent had a 73% chance of being picked by another [24].
- **Specific needs change the pick.** For a product used daily, the Dyson won 88% of runs. In generic comparisons, the cheaper Shark won 91% [24].
- **Checkout friction matters.** Agents finished checkout in 81% of runs at stores with guest checkout. Every run at a store that required login stopped early [24].
- **Marketplaces capture much of the spend.** Marketplaces and aggregators took nearly a third of observed spend, though travel skews that figure [24].

a16z is explicitly looking for startups that help businesses sell to agents [24].

Agents can generate faster than anyone can check

Volunteers pointed coding agents at Alessandro Masullo's unrefereed 940-page draft proof of the generalized Riemann hypothesis. In about three days they merged 2,314 pull requests. The project then paused: each check took 40–80 minutes while about 50 PRs arrived per hour [25]. The main theorem is still disputed [25].

Replit's swarm research points the same way. Opus 5.5 scaled from a 3.7x speedup with one agent to 18.9x with 32. GPT-6.2 Astra got worse beyond 8 agents. Individual model intelligence poorly predicted how well a swarm performed [26]. For investors, the bottleneck is shifting from generating work to verifying and coordinating it.

Sources

1. X post by @nvidianewsroom
2. r/artificial post by u/norcalnatv
3. X post by @alliekmiller
4. X post by @amasad
5. "Fire the Bankers": Most VCs Won't Survive AI | Why the Inference Layer is Uninvestable | Anj Midha
6. X post by @HarryStebbings
7. X post by @danijar
8. X post by @a16z
9. X post by @nathanbenaich
10. Import AI 482: Huawei kernel benchmarks; airborne drone detection; and a national lab's AI science benchmark
11. X post by @ibab
12. X post by @a16z
13. X post by @colossusmag
14. X post by @DominikTornow

15. X post by @amasad
16. X post by @AravSrinivas
17. X post by @AravSrinivas
18. X post by @AravSrinivas
19. X post by @martin_casado
20. Why OpenAI Acquired a Biotech Startup
21. Sam Altman and Greg Brockman on the Next Next Thing
22. Sam Altman, Dario Amodei & Yoshua Bengio Warn UN Security Council of Catastrophic AI Risks
23. X post by @OpenAI
24. X article by @venturetwins
25. r/artificial post by u/michaelTM_ai
26. X post by @amasad