

Crusoe raises at \$30.9B as Stripe’s OpenRouter deal and cheap decision models shape the agent stack

VC Tech Radar

2026-10-04

Crusoe raises at \$30.9B as Stripe’s OpenRouter deal and cheap decision models shape the agent stack

By VC Tech Radar • October 4, 2026

Crusoe raised at a \$30.9B valuation and made its case on GPU economics. OpenRouter explained why it sold to Stripe. Developers disputed Cloudflare’s decision-model speed claims, and a new White House AI task force leaves risk management mainly to industry.

Crusoe’s \$30.9B round and its case on compute economics

Crusoe raised a \$3.9B Series F at a \$30.9B valuation [1]. Its CEO used a 20VC interview to push back on several common bear arguments about AI infrastructure.

- **The real constraint.** He says the bottleneck is powered sites: “there are not places to plug in GPUs” [1]. Crusoe’s answer is vertical integration. A medium-voltage power distribution center that vendors quoted at 100 weeks was built in-house in 28 weeks, and he says the reason was availability, not margin [1].
- **Depreciation.** Crusoe depreciates GPUs over six years, which it calls the industry standard, and expects managed services to keep chips earning beyond that [1]. He claims rental rates for Hopper GPUs are higher today than when they launched three years ago [1].
- **Contract mix.** Crusoe combines five-year contracts with creditworthy customers, shorter deals that pay more but carry more risk, and managed inference and fine-tuning [1]. GPU rentals are usually take-or-pay [1].

Managed GPU clusters earn the highest margins right now because supply is short [1].

- **What it means for startups.** Customers have to forecast compute needs further ahead. Early-stage startups asked to commit to capacity in 2028 are committing to “an eternity.” Crusoe’s pitch is small, modular, factory-built data centers that can deliver capacity closer to when it is needed [1].

Demand is the open question. Epoch estimates the infrastructure could run hundreds of millions to billions of agents. Even 20% utilization would imply \$2.6–5.3T a year in spending, against roughly \$1T in lab revenue by end-2027 [2].

OpenRouter explains the Stripe deal

In his first podcast since Stripe acquired OpenRouter, co-founder Alex Atallah said he “was not thinking about [selling] at all” and that Stripe was his top choice among potential acquirers. He said OpenRouter keeps control of its brand, roadmap and product and gets “a much more serious go-to-market plan.” His view is that “payments and inference are going to blend together” [3]. a16z’s framing of the episode: enterprises are spreading their spend across labs and open-weight models as boards ask about costs and benchmarks, and Replit is building a layer that runs any model on any cloud [4].

Three ideas from the conversation are worth tracking: - **Decision models as an alignment layer.** OpenRouter has an internal prototype that uses a cheap decision model such as Jev to check every tool call against the agent’s system prompt, and against guidelines the agent never sees. Atallah expects it to be paired with structural safeguards [5]. - **Specialist agents vs. one superagent.** Atallah argues for “10 specialized chiefs of staff” over one superagent, because when each has one job you can see where it fails. Amjad Masad runs a single agent across his whole company [6, 4]. - **Models that train their replacements.** Masad suggests general models could train narrow replacements on the fly, much like a just-in-time compiler. The narrow models would be cheaper, harder to hijack with prompt injection, and less harmful because they are less capable [7].

Users dispute Cloudflare’s decision-model speed claims

Cloudflare’s Clef and Clef-flash are built on Qwen and compete directly with TypeSafe’s Jev. Cloudflare reports median latencies of 39ms and 209ms and says that beats rivals [8]. Two developers reported the opposite. One found Clef twice as slow and 2–5x the cost of Jev [9]. Another said Jev beat Clef on speed, accuracy and price, even though their own app runs on Cloudflare [10].

Jev is already built into developer tooling. LangChain’s ModelRouterMiddleware uses it to pick the model for a run, and it is cheap enough to re-pick

after every tool result [11]. An independent evaluation across 16,379 benchmark requests calls Jev “a smaller, humbler model” than its marketing claims, but “genuinely useful” for a niche no one else serves in quite the same way [12].

Model prices keep falling

- OpenAI priced GPT-6.1 Sol at \$2/\$10 per million input/output tokens, against \$10/\$50 for Astra [2]. On Agent Arena it costs 81% less per task than Astra and scores within 1.04 points of it. Anthropic models hold the top three spots [2].
- Bindu Reddy claims Gemini 4.0 Pro is 5x cheaper than Astra at about 95% of its performance [13].
- Aleph Alpha released Kolibri: 78B parameters (3.46B active), up to 1M tokens of context, open weights under Apache 2.0 [14].

Early-stage signals

- **Underdog** announced backing from a16z, Khosla, Hummingbird and Anthology (Anthropic/Menlo), plus angels including Patrick Collison and Naval [15]. Its premise, which it calls “Underdog’s Law,” is that today’s frontier intelligence reaches consumer devices within six months. It designs its models and inference engines together [15].
- **Suhail’s** unnamed startup posted two “biggest revenue day” milestones in a row [16, 17]. Earlier in his thread: a closed seed round, 64 B300s, then “much greater quantities of compute” [18, 19, 20].
- **Perplexity** wants to own its agent sandboxes and will begin rolling out Perplexity Computer on NVIDIA’s Vera CPU, which Aravind Srinivas calls “far better than x86” [21].
- **Google’s Project Suncatcher** prototype satellite, built with Planet, is in orbit and operating as expected. It will test how TPUs hold up to radiation and thermal extremes in space [22].

Agents and incumbents

- **Amazon vs. personal agents.** An investor and board member at Instinct, who also backed the team behind Muse, argues Amazon is stuck. Supporting horizontal agents puts \$80B a year of ad revenue at risk. Resisting them opens the door for Walmart and Shopify [23]. The author proposes a “Prime+” tier at about \$299 a year that includes agentic purchasing [23].
- **Okta’s re-rating.** The stock tripled in five months while revenue grew 11%. The forward earnings multiple went from about 18x to about 50x on cRPO growth reaching 14% and new products making up 30% of bookings [24]. Okta guides Q3 cRPO growth to 11–12%. SaaStr says a print of 11% would remove the main support for the higher valuation [24].

- **AI pricing at Chargebee.** Chargebee says every AI company it talks to has changed pricing at least twice, typically from seats to credits to actions to outcomes [25]. CodeRabbit charges per active agent minute. Gorgias splits pricing between ticket volume and resolved conversations [25].

Policy and safety

A new White House “Super Intelligence Force,” chaired by DNI Jay Clayton, has 120 days to report on AI’s risks and the federal government’s role [26]. Industry stays the primary vehicle for managing risk. The task force’s charter aims to prevent “overregulation and regulatory capture” [26]. Vice chairs include OPM Director Scott Kuper and FTC Chairman Andrew Ferguson. David Sacks joins as an external participant [26].

The OpenAI staffer who led the safety reports for each major launch resigned, arguing the core problem is culture rather than rules [27]. At the UN Security Council, Hugging Face’s Clément Delangue said closed frontier APIs blocked his team from defending against an autonomous agent cyberattack, so they used the open model GLM 5.2. He called for mandatory sharing of full agent traces [28].

Sources

1. Crusoe CEO: Why Everyone Gets GPU Depreciation & AI Energy Costs Wrong
2. [AINews] not much happened today
3. X post by @a16z
4. X post by @a16z
5. X post by @a16z
6. X post by @a16z
7. X post by @a16z
8. r/Futurology post by u/Gari_305
9. r/Futurology comment by u/tribat
10. r/Futurology comment by u/tribat
11. X post by @hwchase17
12. r/MachineLearning post by u/enn_nafnlaus
13. X post by @bindureddy
14. X post by @Aleph__Alpha
15. X post by @0xSigil
16. X post by @Suhail
17. X post by @Suhail
18. X post by @Suhail
19. X post by @Suhail
20. X post by @Suhail

21. X post by @AravSrinivas
22. r/Futurology comment by u/Gari_305
23. X article by @mvernal
24. Why Okta Tripled on ... 11% Growth. 5 Interesting Learnings From One of the Biggest Re-Ratings in B2B This Year: 14% cRPO Growth, 30% of Bookings From New Products, and ~50x Forward Earnings
25. SaaSr AI App of the Week: Chargebee. The Billing System Behind How Gorgias, CodeRabbit, Lambda and Zapier Price AI
26. r/Futurology post by u/John3262005
27. r/Futurology comment by u/Gari_305
28. "OPEN-SOURCE AI IS OUR ONLY DEFENSE!" Hugging Face CEO Clément Delangue Briefs UN Security Council