Inference money keeps flowing (Modal, Applied Compute) as weight-space programming appears in frontier models

VC Tech Radar

2026-11-03

Inference money keeps flowing (Modal, Applied Compute) as weight-space programming appears in frontier models

By VC Tech Radar • November 3, 2026

Large rounds for inference and post-training companies, an early but reproduced result on LLMs editing model weights directly, finance firms restricting coding agents, and robotics, space and defense hardware rounds.

Inference and post-training draw big rounds, with GPU spending weighing on cash flow

Modal is the clearest data point. On 20VC, founder Erik Bernhardsson confirmed $1B in revenue and said the company grew 50x in the last 12 months. Growth is still running 30–40% a month [1]. Harry Stebbings puts sandbox revenue alone at $300M. He calls Modal a $15.75B company and predicts "3-4 specialised inference providers worth $300-$500BN" [2]. Two points from Bernhardsson matter for underwriting:

- **Margins:** selling inference on its own is "a commodity business with no switching costs." What keeps customers is a broader set of products: post-training, sandboxes, agentic tooling and RL [2]. He expects long-term gross margins of 40–60%, not software-style 90%, because every layer of the stack takes 20–30% [2].
- **Capital:** Modal says it is profitable on a P&L basis, but "cash flow is quite negative" because it is buying GPUs and building capacity [1]. He hopes the newly launched training product can reach a $100M run rate within 6–12 months [1].

Applied Compute, a post-training company, raised $350M at a $3.25B valuation, led by Elad Gil. It reports a $100M annualized run rate, reached from zero in 18 months. Its inference service handles 1T tokens a day, 12 weeks after launch. 40% of customers use it for both custom training and inference [3]. Named customers include Microsoft, Shopify, Harvey, Cognition and Decagon [3]. Both companies are pairing training with serving, which fits Bernhardsson's view that breadth of product is what retains customers.

Startups are also building markets for compute itself. **Ornn**, which routes short-term GPU capacity, is reportedly raising a triple-tranche round at a top valuation of about $1B, with Thrive and General Catalyst in talks. It raised a $33M seed led by a16z crypto in June [4]. Silicon Data runs a GPU price index, and Liquid Compute has applied to launch a US compute futures exchange [4]. None of these has regulatory approval yet. The CFTC is asking whether compute is fungible enough to trade as futures, and two investors told Newcomer they passed because the shortage may be temporary [4].

On the negative side, the FT and WSJ report (citing unnamed sources) that OpenAI cut its fundraising target by $10B. It has about $11B in commitments at a $1.25T valuation, against an original goal of $30B at $1.4T [5]. Further down the stack, Marvell agreed to buy Lumentum, a maker of optical components, for $115B in stock. Per Bloomberg, it is the year's largest M&A deal by equity value [6].

Blacksmith (YC W24), which builds CI infrastructure, raised about $66M at a $1B valuation, led by Sequoia, after 20x growth. Its pitch is that agentic coding has made validating code, rather than writing it, the bottleneck [7].

Weight-space programming: an early capability that has been reproduced

Rikiya Takehi reports that GPT-6 Astra edited 0.016% of Qwen3.5-0.8B's weights directly, with no gradients or training. He says this fixed reasoning loops and improved math by up to 6x and instruction following by 3x [8]. An independent researcher says he reproduced the results and that the capability appears only in the newest set of models [9]. Martin Casado asked whether it might be an artifact of labs' recursive self-improvement efforts. That is speculation, not an explanation [10].

The claims are narrower than the headlines. In the game tests, Astra identified networks from weights that carried recognizable spatial clues, and it was told what the input channels meant. It failed to detect sycophancy from weights alone. Its hand-written controllers beat 100M-step PPO in all three games but beat 500M-step PPO in only one [11]. This is worth tracking for interpretability and model-repair startups. It does not yet show that weight editing can replace training.

On agent capability more broadly, Realtime-Bench found the best model, GPT-

6 Astra (high), passed only 36% of tasks under real-world latency. With the clock paused, models reach about 70–80% [12]. Mistral released ML4 as open weights with a technical report, and claims frontier performance on code agents and cyber [13].

Finance firms restrict coding agents; security tooling follows

A widely shared post says Goldman Sachs banned bankers and traders from AI tools such as Claude Code [14]. Amjad Masad called this a sign that agents on employee endpoints are a "systemic security risk" [15]. Replit's advisory argues managed, isolated cloud environments should be the default [16]. It cites a Salt Security survey: 86.4% of respondents had at least one agent running, but only 17.5% could automatically stop one that misbehaved in production [16]. Replit also admits that its sources do not show different breach rates for local and cloud agents [16]. Replit sells cloud agents, so it benefits from this argument.

Two incidents support the case, both reported by parties with something to sell:

- A security vendor says Cursor's marketplace was hacked on Oct. 29. It says a backdoored Claude Code extension auto-installed for 11 hours and left SSH keys and reverse tunnels behind after the fix [17].
- A Reddit post says IBM disclosed two unauthenticated Langflow flaws, an auth bypass and remote code execution, affecting versions 1.0.0 through 1.12 [18].

Agent isolation, credential brokering and auditing look like an active buying category.

Hardware rounds: robotics, space, defense

- **Tutor Intelligence** raised a $75M Series B led by a16z, with USV and Neo. It is in almost 4x as many sites as a year ago and expects more than 1,000 robots under contract by end-2027 [19]. Robots are billed by the hour with no upfront cost. Human "tutors" take over when the AI gets stuck, which generates training data [20]. The company reports 40–60 days from first call to signed contract and about 40 more days to install. Almost all buyers have never bought a robot before [21]. The founder says ongoing costs (inference, teleoperation) now outweigh upfront cost [21]. That makes the cost of tutors falling over time the key metric to watch.
- **Northwood Space**, which builds Earth-to-space connectivity, announced a $375M Series C, under three years after launching as a six-person team [22]. It consists of $300M equity and $75M debt from HSBC Innovation. The company has deployed two product lines and tripled headcount this year [23].

- **Hypersonica**, based in Munich and London, raised a €100M Series B co-led by Lightspeed and NGC to industrialize production [24]. It reportedly went from blank sheet to a Mach 6 test flight in nine months on less than 25M raised [24]. The thesis rests on Europe having no sovereign ground-launched long-range strike system [24].

Other signals

- **Pricing for AI apps:** a16z argues an app is exposed when customers can reproduce its core output with a general-purpose model [25]. In its August survey of 100 enterprise buyers, 94% expected LLM spending to rise and 37% expected it to at least double [25].
- **Qwen gets Apple distribution:** Apple Intelligence launches in mainland China in December with iOS 27.3, using Alibaba's Qwen instead of ChatGPT. It is a partial launch without the new Siri AI [26].
- **Data-center politics:** per Pew, 57% of October political ads mentioning data centers opposed them and only 3% supported them [27].

Sources

1. Modal CEO: Has Open-Source LLM Inference Peaked? | What Happens to the Cost of Compute
2. X post by @HarryStebbings
3. X post by @ypatil125
4. Compute Crunch? Startups Aim to Financialize It
5. r/artificial comment by u/LinkedInNews
6. r/artificial comment by u/LinkedInNews
7. The $1B Company Fixing AI's Software Bottleneck
8. X post by @rikiyatakehi
9. X post by @AlexKrentsel
10. X post by @martin_casado
11. r/MachineLearning comment by u/Disastrous_Room_927
12. Import AI 479: Spacewar; low-latency tasks versus LLMs; Genie 4
13. X post by @PierreStock
14. X post by @exec_sum
15. X post by @amasad
16. X article by @lhchavez
17. X post by @SaoudKhalifah
18. r/deeplearning post by u/No-Conclusion3720
19. X post by @joshgruenstein
20. X post by @a16z
21. What iPhone Packaging Reveals About the Robotics Opportunity
22. X post by @NorthwoodSpace
23. X post by @bridgitmendler

24. Closing Europe's Defence Gap With Hypersonics | Philipp Kerth, Hypersonica
25. X article by @TugceErten2022
26. r/artificial post by u/winer666
27. Political ads turn against data centers