

DeepSeek's Reported Cost Shock Meets a Fragmented AI Stack

VC Tech Radar

2026-08-02

DeepSeek's Reported Cost Shock Meets a Fragmented AI Stack

By VC Tech Radar • August 2, 2026

The clearest signal is a reported $105\times$ cost advantage for DeepSeek V4 Flash, arriving as frontier labs diversify their compute and take divergent platform positions. The brief also tracks Simile's rapid financing, new technical teams, and the growing importance of verification and infrastructure execution.

1. Funding & Deals

Simile raised \$200 million in an unusually fast, insider-led round, taking total funding to \$300 million in roughly six months. Founder Jun Sung Park said the company had raised \$100 million about five months earlier, was not running a process, and was preempted after insiders saw unusual traction, technical progress, and a need for more compute. Green Oaks joined after tracking the market and moved within days; Park identified Index's Shardul as the prior-round lead and Mike Volpi and Astar among the seed backers. [1]

The thesis is not a conventional frontier language model: Simile describes a foundation model of human behavior for simulating individuals, subpopulations, and eventually markets. It wants models that reproduce human mistakes, biases, values, and preferences, using transaction and observational data alongside randomized trials and A/B tests to model causal mechanisms and counterfactuals. [1] Park says published validation predicted behavior and attitudes 85% as accurately as people reproduce their own, while enterprise customers have closed in about three months and used Simile to reproduce findings from three-to-six-month studies in two minutes. [1] The diligence question is whether that data-and-causal-model loop can cross the company's own proof-of-concept-to-production chasm. [1]

2. Emerging Teams

Suhail’s new venture is showing both frontier ambition and infrastructure fragility. The build log records a completed seed round, validation of a basic RLVR post-training stack, a first hire, and a search for a second hire in post-training or low-level model optimization. [2, 3, 4] The project began with two 8xB200 systems and later acquired 64 B300s; in the latest update, Suhail said a key research component was working but that all GPUs had been lost and scaling was delayed by networking problems. [5, 6, 7] For an investor, systems reliability and access to usable compute are part of the research execution risk, not merely an operational footnote.

itnetic is a sharp early security-infrastructure wedge from a one-person team. A Czech solo developer built a Rust reverse proxy for low-rate, human-like Layer-7 attacks that evaded ordinary volume-based defenses, combining JA4+ TLS fingerprinting, half-space-tree/EWMA anomaly detection, and CDN caching. The product is live with a free tier and has handled an attack of 100,000 requests per second. [8] The signal is not revenue yet; it is a narrowly defined operational pain, a technically differentiated implementation, and evidence of deployment under real attack conditions.

3. AI & Tech Breakthroughs

OpenAI says an internal version of Astra produced ten results on long-standing problems in mathematics and theoretical computer science. The company lists advances spanning sphere packing, coding theory, group theory, quantum games, lattice cryptography, Ramsey numbers, and extremal graph theory. It says the total discovery-token cost would have been roughly \$2,000 at Sol API rates; humans prepared the manuscripts, and the model formalized each argument in Lean certificates. [9] This is a significant capability signal, but the investable question is whether independent mathematicians can reproduce and extend the work: OpenAI itself says it takes responsibility for correctness while asking the mathematical community to engage with the results. [9]

DeepSeek V4 Flash is turning the cost-performance story into a deployment story, though the headline remains contested. An analysis cited by @kimmonismus reports that it completes the same benchmark tasks as Fable 5 at 105× lower total cost; Perplexity CEO Aravind Srinivas called two-orders-of-magnitude improvements rare and significant. [10, 11] A community post lists \$0.09 input and \$0.18 output per million tokens with a one-million-token context, while community recipes report serving the 284-billion-parameter model on one DGX Spark at 1,000 tok/s prefill and 59 tok/s in multi-agent serving; a two-Spark FP8 setup reports 82 tok/s single-stream. [12, 13, 14] The counter-signal matters: Bindu Reddy calls the model “benchmark maxx’d,” and another commenter rejects the comparison with Opus 4.8. [15, 16] Treat the 105× figure as a reported benchmark-cost result, not yet as settled capability

equivalence.

4. Market Signals

The frontier compute stack is diversifying away from Nvidia. Nathan Benaich’s refreshed State of AI compute index, with a cutoff of August 1, says Anthropic added up to 2 GW of AMD MI450s, making non-Nvidia silicon 7 of its 8 GW of contracted compute; it puts OpenAI’s non-Nvidia share at 16.75 of 26.75 GW across AMD, Broadcom, and Cerebras. [17] The implication is not that Nvidia has been displaced, but that accelerator mix, software compatibility, and supply access are becoming first-order diligence variables for model and infrastructure companies.

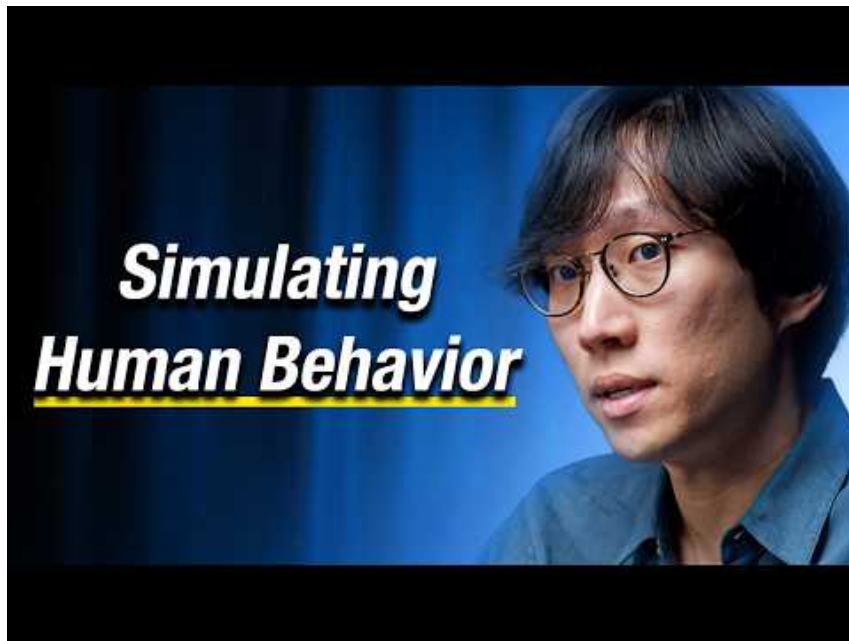
Platform strategy is splitting between “intelligence as a utility” and vertical integration. Garry Tan describes OpenAI’s current direction as an open platform offering intelligence on tap, while a separate post says Anthropic has been telling CEOs, VCs, and startups that it does not see the model and the application or harness as separate companies—and will therefore compete with its customers. [18, 19] These are operator interpretations rather than formal strategy documents, but they give application investors a concrete set of questions: how portable is the product across models, and when does the model vendor become the most dangerous competitor?

AI adoption may require a longer learning horizon than the financing cycle. Exponential View models three companies with the same starting economics and a 5% hit rate but different learning practices: after two years all are losing similar amounts, the eventual loser looks best in year five, and it takes eight years to see which approach produces outsized ROI. The same issue describes a \$45 billion, roughly four-times-levered AI-capex fund that was forced to liquidate after the Philadelphia Semiconductor Index fell 28.6% from its June peak, while noting that the unwind does not prove the underlying thesis wrong. [20]

A separate investor-sentiment signal is emerging in robotics: one post says funds are rewriting 2024 humanoid theses toward vertical-specific solutions, and Bain Capital Ventures’ Ajay Agarwal endorsed it with “Yup.” [21, 22]

5. Worth Your Time

- **Simile interview: the causal-data thesis.** Park explains why the company wants models that reproduce human behavior rather than optimize for super-rational intelligence, and why transaction data, experiments, and counterfactuals matter. [1]



The Best AI Companies Have Unique Data Acquisition Strategies | Simile Co-founder & CEO (6:36)

- **The agent-artifact thread.** A builder says the bottleneck in AI-assisted engineering is preserving intent, specifications, provenance, review, and knowledge transfer—not another context-window increase. The proposed durable unit is an artifact with an owner, version, and acceptance test, reinforced by human approval and diff review before changes are committed. [23, 24, 25, 26]
- **Karpathy’s Opus 5 world-building experiment.** With a roughly \$10, one-million-token budget, Opus spent about two hours writing 5,500 lines of Three.js to render a procedural Lord of the Rings scene; the same experiment exposes the remaining weakness in multimodal self-audit, because the model could not efficiently perceive or play-test the world it created. [27]

Sources

1. The Best AI Companies Have Unique Data Acquisition Strategies | Simile Co-founder & CEO
2. X post by @Suhail
3. X post by @Suhail
4. X post by @Suhail
5. X post by @Suhail

6. X post by @Suhail
7. X post by @Suhail
8. r/SideProject post by u/AdministrativeBat665
9. Ten advances in mathematics and theoretical computer science | OpenAI
10. X post by @kimmonismus
11. X post by @AravSrinivas
12. r/Futurology comment by u/BocciaChoc
13. X post by @bleysg
14. X post by @MiaAI_lab
15. X post by @bindureddy
16. r/Futurology comment by u/ShadowBannedAugustus
17. X post by @nathanbenaich
18. X post by @garrytan
19. X post by @chetanp
20. Leopold & exponential markets; transformative GLP-1s; runaway AI & the future of safety++ #595
21. X post by @antopatrex1
22. X post by @ajay_bcv
23. r/startups post by u/rodrigopfraga
24. r/startups comment by u/CODE_HEIST
25. r/startups comment by u/xJSHAx
26. r/startups comment by u/Visible_Speed8843
27. X post by @karpathy