Seed Conviction Rises as Agent Infrastructure and Deep-Tech Teams Emerge

VC Tech Radar

2026-05-13

Seed Conviction Rises as Agent Infrastructure and Deep-Tech Teams Emerge

By VC Tech Radar • May 13, 2026

A* III and a16z speedrun sharpen the early-stage funding picture while YC and indie founders surface in data-center cooling, agent software, biotech, and on-device AI. Enterprise buyers are moving multi-model, inference costs keep falling, and workflow connectivity is emerging as the next control layer.

1) Funding & Deals

- **A* III** announced a **$450M** early-stage fund. Kevin Hartz's thesis is straightforward: be a founder's first believer, invest before consensus, traction, or even a product, then concentrate time and capital from day zero. A* said that approach has already produced early positions in Ramp, Decagon, Whop, Cape, Simile, Paraform, Watney Robotics, and Mercor, and that the firm now manages **$1B+ AUM** less than five years after launch. [1]
- **a16z speedrun 007** is another strong signal that elite capital keeps moving earlier. The program says it will invest up to **$1M** in brand-new startups, including some that are pre-launch, pre-traction, or even pre-idea, and pair that with **$5M** in tool credits plus operator support across recruiting, GTM, marketing, HR, and more. Andrew Chen's rationale is that AI has compressed team size, timelines, distribution, and iteration speed, making it unusually good timing for small teams to start sooner. Applications close **May 17** for the **July 27–October 11** San Francisco cohort. [2]
- **Isomorphic Labs** raised **$2.1B** to accelerate AI drug discovery. The company tied the raise to its AlphaFold lineage and its mission to reimagine drug discovery. For early-stage investors, this is less a seed datapoint than a category-level validation signal for AI-native therapeutics. [3]

2) Emerging Teams

YC launches

- **InstaAgent** has the clearest traction signal in the YC batch: YC says the company helps B2C brands scale social media marketing across hundreds of personas and has already reached **$1M ARR in 10 months**. Founders: @klwongkyle and @tseungcolin. [4]
- **Madrone** is a hard-tech team worth screening against the AI infrastructure buildout. YC says its dew-point cooling systems can cut power and water use by **30%** at Texas data-center sites. Founders: @akshaytree and @ErikMeike. [5]
- **Superlog** is a sharp devtools wedge: a wizard configures logs, traces, alerts, and dashboards daily, while an agent investigates incidents and posts one mergeable PR per issue into Slack. Founders: @nicolomagnante and @arseniycodes. [6]
- **FinalDose** is one of the more ambitious biotech launches: YC describes it as a programmable drug platform built around a smart molecule that finds diseased cells by DNA and destroys them, starting with cancers. Founders: @Jeffliu6068Liu, @sklin_lite, and @liyaohuang2. [7]

Indie watchlist

- **Alt** pairs a clear privacy wedge with serious technical execution. The KAIST student team says the note-taking app runs fully on-device, offers unlimited free transcription, uses a quantized **1.6GB** voice model for Apple Silicon, hits **12ms** per audio chunk versus a **46ms** benchmark, and runs local Pyannote diarization. It works offline on M-series Macs, iPhones, and iPads, with an optional **$4/month** tier for cloud summaries and translations. [8]
- **firsteyes AI** shows an early lean-distribution signal: six weeks after launch, the solo founder reports **900+ visitors**, **100+ signups**, **250+ audits**, and small revenue with **zero paid ads**, though the landing page still creates confusion for some visitors about the core value proposition. [9, 10]

3) AI & Tech Breakthroughs

- **Perplexity's GB200 work** is the clearest infrastructure update in the batch. The team published how it serves post-trained Qwen3 **235B** MoE models on NVIDIA **GB200 NVL72 Blackwell** racks, arguing GB200 is a major step up over Hopper for high-throughput inference on large MoEs and that it changes how prefill/decode disaggregation should be done. [11, 12]
- **LlamaIndex's liteparse-server** turns document parsing into self-hostable infrastructure: an open-source HTTP server that parses PDFs, Office files, and images and generates screenshots while staying **100%**

self-hosted, private by default, and built for production. The underlying LiteParse parser is model-free, handles **50+** document types, parses complex layouts and tables in seconds, and includes lightweight OCR. [13, 14]

- **The deployment surface around liteparse-server is mature enough to matter.** It ships as a Docker container or serverless Express API and integrates with Redis, OpenTelemetry, Jaeger, Prometheus, and Grafana, which makes it easier to drop into production document workflows. [13]
- **Hugging Face crossing 1,000,000 public datasets** is a quieter but important platform shift. The dataset base doubled in the last **8 months** after taking **4 years** to reach the first **500,000**, and Clement Delangue says better data is becoming the next bottleneck for builders who want to train models themselves rather than just use APIs. [15]
- **Open-source inference optimization is still moving fast under the surface.** Bindu Reddy points to DeepSeek v4 using SSDs for KV cache plus TurboQuant and Kimi K2 compressing memory, arguing that constraint-driven open-source teams are attacking the KV-cache bottleneck directly and pushing intelligence costs down. [16]

4) Market Signals

- **Enterprise AI is becoming a multi-model market.** In SaaStr's cited adoption data, OpenAI remains #1 at **56%**, but Claude has climbed to **48%** and Gemini to **40%**. The article's broader takeaway is that single-model architectures are becoming a procurement liability, coding assistants are driving the fastest revenue growth, and existing distribution contracts still matter enormously. [17]
- **Cheaper inference is shifting value from model access to product design and post-training.** Sarah Guo argues same-task inference costs should fall by at least an order of magnitude per year, which changes the kind of UX startups can build and makes product experimentation more important than training. Her 2026 prediction is that long-horizon agents will push many domain-focused AI companies into post-training, with coding as the first clear example; Clement Delangue agreed. [18, 19, 20]
- **Agent infrastructure is converging on the workflow-connection layer.** A recurring founder view is that the next breakout products will connect agents to the systems people already use—not more wrappers, but workflow access via MCP and similar plumbing. Shopify, Xero, QuickBooks, Figma, and Linear already have versions of this, while sectors such as healthcare, legal, real estate, field service, and education still lack it. [21]
- **Production MCP is harder than the demos imply.** Truto says static OpenAPI-to-MCP generation broke in production because enterprise customers needed custom fields, parameter formats, permissions, and

LLM instructions, forcing environment-level overrides and dynamic tool generation. That is a useful diligence check for anyone evaluating agent-connectivity startups. [22, 23, 24]

- **Monetization and distribution remain the main filter.** Elizabeth Yin notes that many people have already built their own AI tools and that the market is currently in a subsidized phase where services can feel nearly free [25]. Founder reports reinforce that warning: Bloort.ai shut down after **8k visitors**, **200 signups**, **10 installs**, and **$0 revenue** despite heavy outreach [26], while builders in AI image generation describe the category as so saturated that distribution is effectively impossible for commodity products. [27, 28]

5) Worth Your Time

- Andrew Ng on building with AI is worth watching for a grounded view of where value is moving: orchestration layers are making complex agent workflows easier to build, but evals, error analysis, and unstructured-data architecture remain hard; Ng also disclosed small personal investments in LangGraph and LlamaIndex. [29]

Andrew Ng: Building with AI Training Guide (11:31)

- A* III announcement is a compact statement of the current seed-conviction thesis: invest before consensus, traction, or even product, then go deep over time. [1]
- Andrew Chen on speedrun 007 is worth reading for concrete early-stage

terms: up to **$1M** even for some pre-idea teams, plus a view that AI has materially compressed the cost and timing of company formation. [2]

- Truto's MCP architecture guide is a practical read if you are diligencing agent infrastructure; the core lesson is that static MCP generation breaks quickly in real enterprise deployments. [22, 23, 24]
- SaaStr's enterprise AI share essay is the fastest read on where enterprise model share, coding-assistant demand, and distribution advantages are actually moving. [17]

Sources

1. X post by @kevinhartz
2. X post by @andrewchen
3. X post by @demishassabis
4. X post by @ycombinator
5. X post by @ycombinator
6. X post by @ycombinator
7. X post by @ycombinator
8. r/SideProject post by u/Exact_Pen_8973
9. r/SideProject post by u/koustubh18
10. r/SideProject comment by u/koustubh18
11. X post by @perplexity_ai
12. X post by @AravSrinivas
13. X post by @llama_index
14. X post by @jerryjliu0
15. X post by @ClementDelangue
16. X post by @bindureddy
17. Who's Winning Enterprise AI Now: Claude Up 128%, Gemini Up 48%, OpenAI Down 8%, Grok Still A Rounding Error
18. X post by @saranormous
19. X post by @saranormous
20. X post by @ClementDelangue
21. r/SaaS post by u/Leather-Part3037
22. r/SaaS post by u/StealthBeing
23. r/SaaS comment by u/iixfrank
24. r/SaaS comment by u/StealthBeing
25. X post by @dunkhippo33
26. r/EntrepreneurRideAlong post by u/mtsya
27. r/SaaS post by u/Pale_Error_8093
28. r/SaaS comment by u/Bacancyer
29. Andrew Ng: Building with AI Training Guide