Ethos's Series A and the New Infrastructure Stack for AI

VC Tech Radar

2026-05-07

Ethos's Series A and the New Infrastructure Stack for AI

By VC Tech Radar • May 7, 2026

This brief covers Ethos's $22.75M Series A, a cluster of AI infrastructure startups spanning chip design, novel compute, and orbital data centers, and fresh signals around inference economics, managed-agent operating systems, and benchmark-led enterprise adoption.

1) Funding & Deals

- **Ethos — $22.75M Series A led by a16z.** Ethos raised a $22.75M Series A with participation from General Catalyst, XTX Markets, Matt Evantic, and Common Magic [1]. The product uses AI voice agents to capture expertise traditional profiles miss, then matches people into expert calls, research, AI training, fractional work, and full-time roles [2, 1]. The company says 35,000 people are joining weekly and users are making $10,000 per month on the platform [1]. Ben Horowitz said he is "very excited to be working with James and the Ethos team" [3].

  "AI shouldn't replace you. It should make you irreplaceable." [1]

- **VoriHQ — $22M Series B in grocery automation.** Vori says it raised a $22M Series B to make every supermarket in America autonomous, targeting a $1.5T domestic grocery market it describes as still running on Reagan-era technology [4]. Michael Seibel publicly backed the company, saying the team is "doing great work" [5].

2) Emerging Teams

- **Recursive Intelligence.** Anna and Azalia launched Recursive Intelligence around a "recursive self improving loop" between AI and chip design

[6]. Their background includes AlphaChip, the deep-RL system used in multiple generations of Google TPUs, Axion CPUs, Pixel and AV chips, with external adoption by MediaTek [6]. Phase one is faster physical design and verification, including a static timing analysis engine that correlates with commercial tools while running 1,000x faster; later phases move toward a workload-to-GDSII platform and, eventually, vertically integrated chips and models [6]. The team combines ex-Claude, Gemini, and Groq LLM talent with chip-design specialists [6].

- **Unconventional AI.** Naveen Rao—whose background includes a neuroscience PhD, an early AI chip company, MosaicML, and Databricks AI—is building brain-inspired hardware based on nonlinear dynamics and oscillator-style computation rather than conventional matrix math [7]. Rao says the company went from no team in January to a full prototype slated for summer tape-out in six months [7]. The stated target is 3+ orders of magnitude better energy efficiency by pushing closer to physical limits than current AI hardware [7].
- **Flapping Airplanes.** Three months after launch, Flappy is focused on data-efficient AI for under-resourced domains such as robotics, trading, science, and long-tail economic workflows [8]. Its approach mixes proprietary algorithms with low-level GPU systems work, including fine-grained primitives and a custom virtual-machine-style framework for workloads that today's frameworks do not express efficiently [8]. The founding team includes Ben from low-level GPU systems and incubation, Asher from Stanford/Cursor/Mercur, and Aidan Smith from Neuralink [8].
- **Minora AI.** From Plug and Play Uzbekistan, Minora is building a four-agent adtech stack for research, strategy, launch, and optimization across more than 1,400 instruments [9]. The company says it generated more than $580k in revenue and $85k in profit last year, has a $2.1M pipeline, and is targeting the US market; the team previously managed campaigns for brands including Xiaomi, Huawei, Yandex, and UnionPay [9].

3) AI & Tech Breakthroughs

- **XBOW / Expo.** XBOW says its autonomous hacking system found a remote code execution vulnerability in Bing Image Search using only a URL, at a list-price cost of $3,000 [10]. The company also says the same black-box-testing system reached #1 on HackerOne globally [10]. One technical detail: the system uses "model alloys," mixing Sonnet 4.0 and Gemini 2.5 so the models compensate for each other's mistakes like pair programming [10].
- **DeepSeek4.** DeepSeek4 pushes the long-context/cost frontier with a 1 million token context window and a Pro model that reportedly needs about 3x less compute than its predecessor; the lighter Flash model needs about 10x less compute [11]. Its core technique is aggressive KV-cache

compression—token-level compression, 128.1x compressed attention, and compressed sparse attention—which the video says cuts KV memory requirements by about 90% [11]. It also reports better long-context retrieval than Gemini 3.3.1 Pro, strong coding performance, and inference pricing 8-30x below Claude, but the tradeoffs matter: it is text-only, the training stabilizers are not fully understood, and quality drops near the context limit [11].

- **Star Cloud.** Star Cloud One deployed five Nvidia GPUs, including an H100, and the company says it was the first to train nanoGPT in space, run Gemini there, and perform high-powered SAR inference in orbit [12]. The larger plan is an FCC-filed 88,000-satellite constellation with about 20GW of inference capacity, dawn-dusk orbit for continuous solar power, and sub-50ms latency to Earth [12]. The founder argues the economic crossover versus terrestrial solar arrives around $500/kg launch cost, versus Starship's designed $10-20/kg [12].
- **GENE-26.5.** The system combines a robotics-native foundation model, a 1:1 human-like hand, a noninvasive glove for motion, force, and touch capture, and a simulator that reduces experiments from weeks to minutes [13]. It is trained across language, vision, proprioception, tactile, and action, and the company says it can execute fully autonomous tasks at 1x speed with one model and one set of weights [13]. Vinod Khosla called the demo reel—robots cracking eggs, slicing tomatoes, and cooking an omelette—"pretty unbelievable for 2026" [14].

4) Market Signals

- **Inference has become its own product category.** One market observer says AI inference platforms grew as businesses shifted to cheaper models to control exploding token budgets, while web deployment remains a strong market for companies such as Vercel, Netlify, and Lovable [15]. Sarah Guo amplified Baseten founder Tuhin Srivastava's view that even with abundant compute, inference remains the bottleneck: "if we have all the compute, good luck running inference" [16]. Separately, one founder building an AI-dependent SaaS argues the cheap-AI phase is ending through tighter quotas, more paid tiers, and premium features moving behind enterprise plans, pushing builders toward multi-provider architectures and more deliberate model tiering [17].
- **The agent stack is being framed as operating systems, not just prompts.** Michael Chomsky called an OS for Claude managed agents a "generational opportunity" and said he could list 30-50 companies that would use it instantly [18]. Harrison Chase pointed to `deepagents deploy` as the open-source direction LangChain is pursuing [19]. The surrounding language is converging: Alfred Lin says the AI era will optimize software building for "direction and leverage" [20], Garry Tan calls that "just in

time software" [21], and Bindu Reddy sketches the next-generation company as half a dozen people running thousands of agents [22].

- **Benchmarks are becoming enterprise adoption tools.** Harvey released LAB, described as the first long-horizon, open-source legal agent benchmark, to help legal teams understand what legal agents can do now, plan deployment, and design human-agent cooperation [23]. Parag Agrawal argued the market needs more long-horizon, real-world work benchmarks and said he is excited to help add the web to one such environment [24].
- **PDF/document parsing is being positioned as agent infrastructure.** Jerry Liu argues AI agents will automate large amounts of knowledge work, but much of the relevant data lives in documents and PDFs that existing OCR tools, frontier VLMs, and current benchmarks still handle poorly, especially around tables and layout [25]. He argues agents need PDF tools both at ingest time and as runtime tools, and positions LlamaParse, LiteParse, and ParseBench as that stack [25].
- **The labor signal cuts against the "job apocalypse" narrative.** a16z's David George points to rising demand for software engineers, software developers increasing as a share of new jobs, above-trend wage growth in AI-exposed industries, and open PM jobs at their highest level since 2022 [26].

5) Worth Your Time

- **Recursive Intelligence at Sequoia.** Covers AlphaChip's lineage, Recursive's 1,000x static timing analysis engine, and the company's workload-

to-GDSII roadmap. Watch here [6] *AI That Designs Its Own Chips: Ricursive's Anna Goldie and Azalia Mirhoseini (5:18)*

- **XBOW at Sequoia.** Covers the Bing Image Search RCE, fully autonomous black-box testing on HackerOne, and the company's "model al-

loys" approach. Watch here [10] *Inside the Rise of Autonomous AI Hackers: XBOW's Oege de Moor*

(1:38)

- **Unconventional AI at Sequoia.** Covers nonlinear-dynamics compute, prototype tape-out timing, and the company's energy-efficiency thesis. Watch here [7]
- **LlamaIndex's PDF-processing deck.** Explains why PDFs remain hard for agents, and how LlamaParse, LiteParse, and ParseBench fit together. Slides [25]
- **Kyber inside Lex Fridman's FFmpeg episode.** Covers 4ms glass-to-glass streaming, QUIC/UDP transport, multi-stream sync, and remote control of robots and drones. Episode [27]

Sources

1. X post by @DJ_Mankowitz
2. X post by @a16z
3. X post by @bhorowitz
4. X post by @IamBrandonHill
5. X post by @mwseibel
6. AI That Designs Its Own Chips: Ricursive's Anna Goldie and Azalia Mirhoseini
7. Why the Brain Computes 1,000,000x More Efficiently Than A GPU: Unconventional AI's Naveen Rao
8. Why Data Is the Real AI Bottleneck: Flapping Airplanes' Ben and Asher Spector
9. Plug and Play Uzbekistan Expo 2026
10. Inside the Rise of Autonomous AI Hackers: XBOW's Oege de Moor
11. DeepSeek V4 AI Beats Billion Dollar Systems…For Free
12. Starcloud's Philip Johnston: Why the Cheapest Compute Will Be in Space
13. X post by @gs_ai_
14. X post by @vkhosla
15. X post by @arakharazian
16. X post by @saranormous
17. r/SaaS post by u/LucianoMGuido
18. X post by @michael_chomsky
19. X post by @hwchase17
20. X post by @Alfred_Lin
21. X post by @garrytan
22. X post by @bindureddy
23. X post by @saranormous
24. X post by @paraga
25. X post by @jerryjliu0
26. X post by @a16z

27. FFmpeg: The Incredible Technology Behind Video on the Internet | Lex Fridman Podcast #496