Pit's $16M Round, Seed IQ's Benchmark Signal, and the New Inference Math

VC Tech Radar

2026-05-08

Pit's $16M Round, Seed IQ's Benchmark Signal, and the New Inference Math

By VC Tech Radar • May 8, 2026

Pit's a16z-led financing was the clearest deal signal, but the broader pattern is a thickening stack for production AI agents across testing, authorization, data access, and security. The macro backdrop is equally important: inference is getting cheaper while total demand, China competition, and engagement-led B2B adoption keep rising.

1) Funding & Deals

- **Pit — $16M led by a16z** [1]. Pit launched with $16M in funding led by a16z, with additional backing from founders and operators across OpenAI, Anthropic, Google, Revolut, Deel, and others [1]. The founders say the product comes from operational pain they saw at Voi, Klarna, and Zettle, and frame the thesis as replacing rigid software with AI-native systems built around real workflows [1]. The key signal is early enterprise pull: Ben Horowitz said a16z is already seeing large enterprises replace manual operational work with Pit, moving faster and freeing teams for higher-leverage work [2, 1].
- **Refactor 5 — $50M for seed hard tech** [3]. Refactor Capital announced Refactor 5, a new $50M fund backing seed-stage hard tech founders across aerospace, bio, critical materials, energy, and robotics, with most portfolio companies expected to have AI at the core [3]. The firm says it has launched five funds over ten years and manages roughly $300M AUM, with Refactor 5 coming online next year while investments continue from Refactor 4 [3]. For investors, this is a useful capital-formation signal around physical AI and AI-native hard tech [3].

2) Emerging Teams

- **Prototyping.io — autonomous manufacturing with real revenue** [4]. Prototyping.io says its systems turn CAD designs into high-quality mechanical parts in as fast as one day, cutting weeks out of hardware iteration cycles for multi-billion-dollar customers while already doing $400k in monthly revenue [4].
- **A practical agent-control stack is forming** [5, 6, 7]. Chronicle Labs is building a staging environment that replays production events in sandboxes so enterprise teams can backtest agents before live deployment [5]. Clawvisor is attacking the authorization layer, letting agents access apps like Gmail and Slack without sharing credentials; users approve tasks once and Clawvisor enforces them [6]. Garry Tan called Clawvisor an important part of making the agent world secure and enterprise-grade [8]. Strukto.ai's Mirage tackles data access by mounting services like S3, Google Drive, Slack, Gmail, GitHub, Linear, Notion, Postgres, MongoDB, and SSH into one versioned virtual filesystem that agents can operate on with standard Unix tools [7]. The execution tempo is also notable: the team says Mirage was built in six weeks with 1.1M+ lines of code [7]. Together these teams map to three practical deployment bottlenecks in production agents: testing, permissions, and data access [5, 6, 7].
- **Dolly — per-employee messaging agents with an early trust signal** [9]. Dolly fine-tunes one agent per employee on that person's communication history to respond to email and Slack with higher voice fidelity than prompt engineering alone, targeting roughly three hours per day of async messaging load [9]. The behavior to watch is user comfort: pilot users reportedly get comfortable delegating routine replies after two to three weeks, and the company is moving from three pilot organizations to a capped group of twenty [9, 10].

3) AI & Tech Breakthroughs

- **Seed IQ and ARC-AGI-3 are the benchmark story to watch** [11]. A post in r/deeplearning says AIX's Seed IQ has an unofficial 100% score on ARC-AGI-3, while top transformer models were below 1% [11]. The same post cites the Arc Prize Foundation's March 25 update to ARC-AGI-3, which replaced static grids with interactive game environments that require active inference and measure skill-acquisition efficiency against humans, who remain the 100% baseline [11]. According to the post, official testing alongside frontier models like Gemini 3.1 may be only weeks away [11].
- **OpenAI is pushing both verticalized and real-time agents** [12, 13]. In security, GPT-5.5-Cyber is rolling out in limited preview to defenders securing critical infrastructure, while GPT-5.5 with Trusted Access for Cyber is positioned as the best option for developers finding and patching

code vulnerabilities [12]. In voice, OpenAI launched GPT-Realtime-2 in the API as its most intelligent voice model yet, alongside GPT-Realtime-Translate and GPT-Realtime-Whisper for real-time voice interfaces [13]. Sam Altman said the cyber push is about helping companies secure themselves quickly [14].

- **VinciPhysics is arguing for a new class of world model** [15, 16]. Hardik Khandelwal and @saucentoss published a paper defining the criteria for foundation models for physics and linking that to continuous physics reasoning [15]. The framing from the team is that physics is the next major world model after language, vision, and code [15]. Vinod Khosla's endorsement translates the thesis into market scale: bringing continuous physics reasoning to 100x more engineers and 1000x more simulations in a fraction of the time [16].
- **Open Design is a strong open-source counter-signal** [17]. Nexu-io's Open Design, positioned as a local-first Apache-2.0 alternative to Claude Design, reached 18k+ GitHub stars in five days [17]. The strongest product wedges are BYOK support across existing AI CLIs, an MCP server that lets editor agents read design artifacts directly, and the ability to draft with cheaper or local models before switching to frontier models for final polish [17].

4) Market Signals

- **Cheaper inference is increasing total compute demand, not reducing it** [18]. The cost of 1M frontier reasoning tokens reportedly fell from roughly $60 to $0.50 in twelve months, about a 128x drop, yet hyperscaler compute bills continue to rise [18]. The explanation is that reasoning models use about 10x more output tokens, agentic workflows chain roughly 20x more requests, and deep-research queries can cost more than 10 original GPT-4 queries, so lower unit costs unlock much larger workloads [18]. Andrew Chen's early Codex /goal usage points in the same direction: he expects unattended 24/7 LLM use to increase token consumption by several orders of magnitude [19].

  > "The math at the aggregate level is brutal: 100x cheaper tokens times 10 000 more tokens equals a 100x larger total bill." [18]

- **China's AI market looks more like cloud than SaaS** [20]. Interconnects' reporting from Chinese labs suggests enterprise AI spend is more likely to track China's large cloud market than its historically smaller SaaS market, with little concern that inference demand will fail to emerge [20]. Two other signals stand out: Chinese developers are reportedly heavily using Claude despite the ban [20], and major incumbents from Meituan to Xiaomi and Ant are building their own general-purpose LLMs to control more of the stack [20]. Nvidia shortages remain acute, while Huawei is viewed as viable for inference [20].

- **In B2B AI, engagement is becoming the leading indicator** [21]. SaaStr argues that DAU/MAU now matters more than ARR growth or NPS because it leads renewal, expansion, and churn in agent-era products [21]. Harvey is the case study: net new ARR up 6x year over year, DAU/MAU nearing 50%, and average users spending 12 hours per month in product [21]. The practical dashboard is DAU/MAU, hours per MAU, queries or actions per MAU, and stealth-churn cohorts rather than just cancellation data [21].
- **Vertical agents are starting to show labor compression** [22]. SaaStr's in-house customer-success agent QBee cut total human hours by roughly 70% across internal and external work while managing more than 150 sponsors, producing a claimed 3x productivity multiplier on the work that remained [22]. The more important product signal is that SaaStr built QBee because it could not find an off-the-shelf AI customer-success agent, and says it would replace QBee immediately if a better third-party product existed [22].
- **Political backlash is still a lagging risk, not a current constraint** [23]. One policy-oriented thread cited survey work showing AI is only Americans' 29th most important issue, arguing that negative sentiment has not yet translated into meaningful political action [23]. The predicted trigger is labor-market pain—roughly a two-point rise in unemployment attributed to AI—with the risk that bad policy responses such as data-center moratoria arrive before better ideas do [23].

5) Worth Your Time

- **My First Million: How Replit made $1M on day one (then $250M in a year).** Best for understanding why agentic coding may expand software demand rather than just compress costs: Replit frames its agent as an early end-to-end breakthrough, says it hit $1M ARR on day one and $2M on day two, and then pivots to the kinds of niche, bootstrapped software businesses that become viable when software gets cheaper to make [24].

How Replit made $1M on day one (then $250M in a year) (18:17)

- **Interconnects: Notes from inside China's AI labs.** Probably the best single read in the set on enterprise demand, open-first model strategy, talent mobility, and compute constraints inside China [20].
- **Demian AI's inference economics thread.** This is the cleanest explanation for why cheaper tokens can still mean bigger bills once agents and deep-research workflows arrive; Nathan Benaich explicitly endorsed the framing [18, 25].

  "The right framing is that AI got dramatically cheaper, dramatically more capable, and dramatically more useful…" [18]

- **Equity Podcast: The long road to driverless with Aurora's Chris Urmson.** Useful for physical-AI investors because Urmson lays out the trucking-before-robotaxis market choice, the First Light lidar unlock for highway safety, and the case for "verifiable AI" over end-to-end opacity [26].

The long road to driverless with Aurora's Chris Urmson (Live at HumanX) | Equity Podcast (24:04)

Sources

1. X post by @mradamjafer
2. X post by @bhorowitz
3. X post by @zalzally
4. X post by @ycombinator
5. X post by @ycombinator
6. X post by @ycombinator
7. X post by @zechengzh
8. X post by @garrytan
9. r/EntrepreneurRideAlong post by u/Substantial-Cost-429
10. r/EntrepreneurRideAlong comment by u/Maya_The_Great
11. r/deeplearning post by u/andsi2asi
12. X post by @fouadmatin
13. X post by @OpenAI
14. X post by @sama
15. X post by @hardikk13
16. X post by @vkhosla
17. r/SideProject post by u/Exact_Pen_8973
18. X post by @demian_ai

19. X post by @andrewchen
20. Notes from inside China's AI labs
21. DAU, WAU and MAU Are the New Lighthouse Metric in B2B + AI. Harvey's a Great Case Study.
22. Our AI VP of Customer Success "QBee" Saved Us 70% of the Human Hours Vs. Before. Both Internally and With External Teams. A 3x Multiplier.
23. X post by @ahall_research
24. How Replit made $1M on day one (then $250M in a year)
25. X post by @nathanbenaich
26. The long road to driverless with Aurora's Chris Urmson (Live at HumanX) | Equity Podcast