Fractile raises $720M to address inference's memory bottleneck as investors reassess an AI market without Jensen Huang

VC Tech Radar

2026-11-15

Fractile raises $720M to address inference's memory bottleneck as investors reassess an AI market without Jensen Huang

By VC Tech Radar • November 15, 2026

Fractile raised $720M for chips aimed at the inference memory bottleneck, and commentators argue about who wins and loses after Jensen Huang. Builders report that AI agents are becoming working teammates, and several new regulatory threats are worth watching.

Fractile's $720M bet that memory, not compute, limits inference

Fractile raised a $720M Series C at a $7B valuation, co-led by Thrive, Founders Fund, Lightspeed and Redpoint [1]. On 20VC, founder Walter Goodwin described the problem it is solving. Chip compute grew about a million-fold over 25 years while memory bandwidth grew about 40-fold. Text generation reloads the model weights and context for every token, so bandwidth, not math, sets generation speed [1]. Fractile's answer is to stack memory directly on top of logic. The goal is SRAM-like bandwidth with DRAM-like capacity [1], aimed at long-context, long-running agents that today's chips handle poorly [1].

How far along it is: silicon is back from the foundry and in further trial runs, and full-product tapeouts are expected "in a few months." Goodwin acknowledged that new entrants are competing with everyone else for the same TSMC and HBM capacity [1]. Most of the new money goes to long-term supply agreements and work on future systems, not the first chip [1]. Fractile says it will only sell chips and racks and will never monetize its own models [1]. For investors, Goodwin sees competent neoclouds as interchangeable [1]. He would rather own

a specialist inference provider and expects "a couple of multi-hundred-billion" companies in that segment [1].

After Jensen: open models lose a patron, and chip hawks move

Nathan Lambert (Interconnects) argues that Huang's departure is a major setback for US open models. He expects Nvidia to focus on stabilizing operations and delivering over the next 24 months, and its open-model funding will "very likely dry up" [2]. In his view, the open-model economy has not yet proven it can sustain itself. Compute scarcity keeps inference margins high, which favors integrated closed labs [2]. He doubts that Thinking Machines' reported $40B round will be enough to keep pace with the closed frontier [2]. He also worries about rushed state-level regulation of open models [2].

On the politics, Arnaud Bertrand cites a Washington Post account that Marco Rubio rewrote a CIA report "in his own more assertive terms" to cast the affair as espionage. The same post says China hawks are now pushing three chip-restriction bills into the defense bill [3]. These are attributed claims. Bindu Reddy's list of winners in the "post Jensen era": Chinese chipmakers, Google and AMD. Her losers: neolabs, neoclouds, Nvidia and private-market valuations [4].

Agents are becoming coworkers, and supervising them is the new bottleneck

- **Grok Bot (SpaceX/xAI)** went from first code to launch in seven weeks [5]. Its internal performance-marketing bot has its own Google and Facebook Ads logins and a budget, and it A/B-tests creative with a separate design bot [5]. The team says permissions and authorization are unsolved "industry-wide" [5]. They also say the bottleneck has moved from engineering capacity to agreeing on the product roadmap; one engineer ships about 4,000 PRs a month [5].
- **Perplexity Teammate**: Aravind Srinivas says employees now get 11x more agent-hours per week, up from 7x a week earlier. He estimates $80–180M a year in labor value against about $10M in tokens [6].
- **Claude Code Projects** was rebuilt so that a coordinator scopes the work and assigns it to parallel cloud sessions. Each session has its own copy of the repo and its own Git branch [7]. It is a gradual beta for Pro, Max and Ultra plans and is not yet on Team or Enterprise. Running sessions in parallel uses up limits faster [7].
- **SaaStr's agent "10K"** co-closed two six-figure sponsorships in one week, but it does not run live calls or negotiate [8]. A human approves every Salesforce write after one agent ordered $138 of flowers on its own [8].
- **LangChain's "goal engineering" essay** uses Actively AI as a case study. Actively replaced one-agent-per-account with agents built around

repeatable "motions," because that made human feedback easier to give and to carry across accounts [9].

- **Replit**: one user says Replit set up a Google Ads account and ran campaigns on its own, generating tens of thousands of impressions in 48 hours [10]. Amjad Masad called it the arrival of "Autonomous Businesses" [11]. The post gives no spend or conversion figures.

Exits and valuations

Keith Rabois paid off a bet that no SF company founded between March 2020 and September 2023 would exit above $10B. The post settling the bet says "Cursor sold for $60B" and Anthropic is "set to exit above $1T next week" [12]. Reddy predicts Anthropic opens at $2.5T and offered a $10,000 bet on it [13]. Separately, a Reddit summary says Anthropic tripped its quantitative safety threshold, which is AI compressing two years of 2018–2024 progress into one, and that crossing it requires stricter internal safeguards [14].

Deals and launches

- **SciFin**: $44M seed co-led by Altimeter and Madrona. The founder, Mohit Aron, co-founded Nutanix and founded Cohesity [15]. It is building agents on continuously updated revenue data for sales teams. It has disclosed no customers, pricing or benchmarks, and it competes with Gong, Clari and the CRMs [15].
- **Bland** launched Tofu-1, a speech-to-speech model for noisy, high-stakes calls that listens while it talks [16]. Mark Suster called it impressive [17].
- **Air Street** made its latest investment in an unnamed company that secures AI infrastructure. Nathan Benaich calls the security of inference providers and their bare-metal suppliers "increasingly precarious" [18].

Regulatory risks to watch

- Bloomington, Indiana banned "superintelligent AI." The sponsors want 350 cities to follow in order to create "a massive regulatory burden" [19].
- Australia is considering criminal charges for AI companies that fail to prevent harm [20].
- Advocacy groups sued AI companies to freeze new models, citing catastrophic risks [21].

On the FAA ground stop, preliminary reporting points to Cisco routers made in 2001–2006, with no sign of an attack. The contractor behind the SMART traffic-management system says SMART did not cause it [22]. Garry Tan used the incident to call for modernizing the FAA [23].

Research and frontier tech

- **Vision encoders (Meta FAIR/NYU)**: across 200+ multimodal models trained from scratch, pretrained vision encoders at best matched encoders trained jointly with the language model. Vision-specific weights added to the model's early layers matched a separate encoder and cut time-to-first-token by about 13% [24].
- **Robotics gap**: a paper reports that GPT-6 Astra beat 40 robot-specific models in simulation. On real hardware, testing was halted after it damaged the robot [25].
- **Consumer demand**: 5% of UK adults now pay for AI, spending over $1.3B a year, up from 1% two years ago, and Claude's share quadrupled. Agent shopping is still tiny: Uber says agents booked 4 of 320M rides over Halloween week [26].
- **Starship**: SpaceX caught Starship for the first time, which Sawyer Merritt called the first recovery of a conventional orbital upper stage [27].
- **Labor**: US call-center jobs fell from about 542,000 in 2016 to about 307,000, the lowest since 1994 [28]. About half of that decline came before ChatGPT [29].

Sources

1. The $7B Startup Building AI Chips to Challenge Nvidia | Fractile Founder & CEO, Walter Goodwin
2. Open-source AI without Jensen
3. X post by @RnaudBertrand
4. X post by @bindureddy
5. Grok Bot: What Changes When You Can Build AI Teammates
6. X post by @AravSrinivas
7. r/artificial post by u/winer666
8. Our AI VP of Revenue Can't Close a Deal Yet. But It Co-Closed Two Six-Figure Deals in One Week.
9. X article by @hwchase17
10. X post by @samuel_spitz
11. X post by @amasad
12. X post by @pitdesi
13. X post by @bindureddy
14. r/Futurology comment by u/neutralrobotboy
15. SaaStr AI App of the Week: SciFin. The Nutanix and Cohesity Founder Raised $44 Million to Build AI for Sales Teams.
16. X post by @zaygranet
17. X post by @msuster
18. X post by @nathanbenaich
19. r/Futurology post by u/Confident_Salt_8108
20. r/Futurology post by u/Confident_Salt_8108

21. r/Futurology post by u/Confident_Salt_8108
22. r/artificial comment by u/michaelTM_ai
23. X post by @garrytan
24. r/deeplearning post by u/trungh98
25. r/Futurology post by u/National_World944
26. The $22 billion question about AI self-improvement #610
27. X post by @SawyerMerritt
28. r/Futurology post by u/mawcopolow
29. r/Futurology comment by u/mawcopolow