

Chip startups draw competitive rounds, cheap "decision models" multiply, and panelists say the traditional seed round is breaking

VC Tech Radar

2026-10-02

Chip startups draw competitive rounds, cheap "decision models" multiply, and panelists say the traditional seed round is breaking

By VC Tech Radar • October 2, 2026

Benchmark wins the lead on Tendrils Compute and Volantis raises \$88M as chip funding continues. Perplexity and Cloudflare open-source small decision models. Investors argue over seed sizing, neolab exits, and whether open-weight models have peaked.

Chip startups keep drawing competitive rounds

Benchmark won a contested bid to lead a new round in **Tendrils Compute**, an early-stage chip startup based in Cambridge, UK, and led by Nils Cremer. People familiar with the deal say Tendrils is already discussing a quick follow-on that could value it above \$1B [1]. The technical bet is unusual: general-purpose chips built on interaction nets, a graph-based computing model. The pitch is that CPU speed becomes a bottleneck as agents chain inference steps and tool calls [1]. Benchmark has done this before. It co-led Cerebras's 2016 Series A and earned about 20.5x when Cerebras went public [1].

Volantis announced an \$88M Series A for optical interconnects that target AI's memory bottleneck. Its claims include inference at up to 10,000 tokens/sec per user on models larger than 10T parameters. It says its next chip iteration has already taped out [2]. Newcomer lists other recent chip rounds: - Etched: \$700M at a \$21B valuation - Fractile: in talks at \$6.5B after an Anthropic supply deal - OLIX: \$312M at \$3.3B - Euclid and Delos Data: nine-figure rounds [1]

There is a counterweight. Cerebras traded as high as \$386 after listing but now sits close to its \$185 IPO price [1].

Cheap “decision models” are becoming their own category

Perplexity is open-sourcing **pplx-decider-27b**, a multimodal model that outputs a probability distribution over a fixed set of answers instead of free text. It is served through a Decisions API at \$0.04 per million input tokens with free output, and Perplexity says the price will drop further [3, 4]. Cloudflare open-sourced its first homegrown decision models, **clef** and **clef-flash**, on Workers AI the same day [5]. Harrison Chase described the use case: cheap typed answers for small calls inside a harness, such as routing, approvals and judging, with a big model handling everything else [6].

Jev from TypeSafe AI started this category. On 20VC x SaaStr, the hosts said TypeSafe raised a \$40M seed at \$200M and is now in talks to raise more than \$1B at a valuation above \$10B. Jason Lemkin cited Jev at 17% of OpenRouter traffic [7]. With two open-weight entrants priced about the same, the main question for anyone underwriting TypeSafe at that valuation is how defensible it is.

LangChain published data supporting routing. In a 973-thread A/B test, its Open SWE harness either routed each task to the cheapest adequate model tier or always used GPT-6 Astra. Merged-PR rates were 29.2% routed and 27.3% control, a difference that was not statistically significant ($p=0.49$). Median cost per thread fell 64%, from \$2.61 to \$0.94 [8]. A fast-model-only arm was stopped within a day because output quality was too low [8].

The 20VC x SaaStr panel on venture now

Benchmark’s Jack Altman joined the panel because Benchmark co-led **Instinct’s** \$1B round at \$10B. That round came 33 days after Instinct was raising at \$2.5B. The founder says Instinct is approaching \$1B in annual transactions, more than half of them travel [7]. Main threads:

- **Neolab exits.** AMD’s \$8.2B all-stock purchase of World Labs, about 2.5 years after founding, was described as the first big neolab exit. One partner’s tally counts 102 neolabs that have raised more than \$70B. Altman’s answer was that about 10 companies can now do \$10B acquisitions, which he called faster than an IPO [7].
- **Seed sizing.** Altman called \$3–6M checks for 8–15% ownership “broken in many lanes.” His reasoning: labs need \$200M or more, and well-networked founders skip straight to \$50M rounds. Lemkin’s view is that \$2–3M is still the natural seed size when investors aren’t competing to fund you [7].
- **Inference and open weights.** Modal tripled its valuation to \$15B and Baseten is in talks at \$26B. Altman called inference an index bet on everything outside the labs [7]. Lemkin thinks open-weight share has peaked. He cites closed-model price cuts and says he couldn’t find an enterprise at Dreamforce willing to run mostly China-origin open weights. Altman expects more enterprise post-training on open weights, but thinks compute

ownership decides the outcome [7].

- **IPOs.** Oura pulled a \$15.62B IPO two days before pricing despite reports the book was oversubscribed. One panelist pointed to Forerunner’s plan to sell its entire stake, which makes sellers very price-sensitive. Altman noted no AI-native application company has gone public yet, and expects several in 2027 if markets hold [7].

SaaStr’s reading of a16z’s market deck adds data on how spread out outcomes have become: - Under-one-year-old B2B companies are growing more than 500% year over year. Mature B2B companies are at about 24% [9]. - 55% of US unicorns have less than two years of runway [9]. - For the 2024 vintage, top-decile VC net IRR is 40.5% against a -3.3% median. Median DPI is 0.00x for every vintage since 2021 [9].

Early-stage rounds and teams

- **doxxnet** (Barrett Lyon, co-founder of Prolexic) raised \$38M led by a16z. The product is private networks for people and their AI agents with no server in the middle, and the company says it has blocked 38M+ threats since December [10, 11]. a16z’s thesis is that agents widen the personal attack surface [11].
- **Halluminate** builds reinforcement-learning environments for knowledge work beyond coding, starting with finance. With fewer than 10 people, it works with four of the top five closed US labs and recently raised a \$30M Series A. Its founders argue verification matters more than task volume [12].
- **Harmonic’s Hot 25** ranks Resolve AI first for the third time. The Latent Co is the top newcomer at #2, with Prime Intellect, Strala AI and Trajectory Labs also new to the list [13].
- **YC** promoted founders Vivian Shen (Juni, Acely) and Raphael Schaad (Cron) to General Partner [14]. Garry Tan says the partnership now has 19 members, the largest in its history [15].
- **Imbue Studio** went to waitlist. It is pitched as a personal computer you shape by describing what you want, open source under FSL, runnable locally, and able to switch between AI providers [16, 17].
- **AMI Labs.** Yann LeCun says it is still in research with no short-term product plan, has 50–60 staff across Paris, New York, Montréal and Singapore, plans more funding rounds to pay for compute, and does not intend to be acquired [18].

Models and research

Google DeepMind’s **Gemini 4 Argon** is available first only to government users and trusted cyber defenders [19]. Artificial Analysis scores it level with GPT-6 Astra, at \$1.99 per task versus \$3.26. The savings come from introductory pricing, not efficiency: Argon uses 62K output tokens per task against Astra’s

27K [19]. Vals lists 262K maximum output tokens against Google’s 1M claim [19].

LlamaIndex says its Extract v2.5 document-extraction agents beat Opus 5.5 and GPT-6 Sol while costing 30% to 4x less [20]. arXiv now limits each submitter to two submissions per calendar month [21].

Sources

1. SCOOP: Benchmark Backs Early-Stage Startup Tendrils Compute as Chip Momentum Builds
2. X post by @semiDL
3. X post by @AravSrinivas
4. X post by @perplexitydevs
5. X post by @ritakozlov
6. X post by @hwchase17
7. 20VC x SaaSr: Anthropic’s S-1 Leaks, Instinct Hits \$10 Billion in 33 Days, AMD Buys World Labs, and Meta Poaches MongoDB’s CEO
8. X article by @sydneyrunkle
9. The 10 Most Important Learnings From a16z’s Latest State of Markets: Horizontal B2B Trades at 2.7x Revenue, New Startups Are Growing 500%+, and 55% of Unicorns Have Under 2 Years of Runway
10. X post by @BarrettLyon
11. X post by @a16z
12. X post by @ycombinator
13. X post by @HarryStebbing
14. X post by @ycombinator
15. X post by @garrytan
16. X post by @kanjun
17. X post by @kanjun
18. [Version longue] Les World Models, un game changer ? Avec Yann LE CUN | AIM 2026
19. [AINews] Gemini 4 Argon: GDM’s answer to Astra/Fable, with 1M output
20. X post by @jerryjliu0
21. r/MachineLearning post by u/Nunki08