

OpenAI DevDay: always-on Dots agents, a Sol
model at a fifth of Astra’s price, and a
marketplace for partners, while Astra 6.1 stays
unreleased

VC Tech Radar

2026-09-30

OpenAI DevDay: always-on Dots agents, a Sol model at a fifth of Astra’s price, and a market- place for partners, while Astra 6.1 stays unre- leased

By VC Tech Radar • September 30, 2026

OpenAI launched always-on agents and cut the price of near-frontier intelligence while holding back its next Astra release. Meanwhile, Hugging Face is being acquired by NVIDIA, General Intuition raised a large round, and AI roll-ups kept growing.

OpenAI DevDay: agents, cheaper intelligence, and Chat- GPT as a sales channel

Dots are always-on agents powered by Astra. Each Dot learns how its user works. It can take ongoing tasks like “monitor incoming bug reports.” In the demo, a Dot migrated an app off a legacy API: it traced the dependencies, wrote code, ran tests and opened a pull request for review [1]. Dots run on their own cloud computer and use plug-ins the user has already connected. Texting and calling a Dot are listed as “coming soon” [1]. ChatGPT Space adds shared pages where teams can tag a Dot to work alongside them [1].

Altman describes Dots as an enterprise-leaning product first, aimed at “the top end of the market,” with a mass-market consumer version to come later. He says it costs more than competing agents because of how capable it is and how much compute it uses [2]. A Bloomberg columnist questioned who the product is for, noting the cheapest plan that includes it is \$100 a month [3]. Allie K.

Miller highlights the dedicated VM behind every Dot, which she sees as a large compute commitment in the same direction Meta took with Muse. She also calls it buggy for now, and sees an early announcement meant to pull attention away from Muse, GrokBot and Instinct [4].

Pricing moved down sharply. OpenAI says 6.1 Sol delivers near-Astra intelligence at one fifth of Astra’s price, with a 95% discount on cache reads. An 8x “ultrafast” mode is available now for Astra and is coming for Sol [5]. Altman’s framing is to “lead the Pareto frontier at every point” on the cost-quality curve [2].

Distribution for developers. OpenAI cites about 1.2B ChatGPT users. Users can now sign into partner apps with ChatGPT and spend the tokens from their ChatGPT plan, so the developer doesn’t carry the inference cost. There are 16 launch partners [1]. Two partners announced marketplace deals: - **Baseten** is among the first open-model inference providers in OpenAI’s B2B Marketplace, with a native Codex integration. Enterprises can spend their existing OpenAI commitments on open models that Baseten serves [6]. - **Runway** joined as a launch partner. Customers can apply part of their OpenAI commitment to Runway purchases [7].

For open-model and specialist-model companies, access to OpenAI’s enterprise commitment dollars is a new sales channel.

Platform pieces: - Codex now runs in the cloud. - A new agent API includes the harness, multi-agent controls and computer use. - OpenAI reports more than 99% reliability and a 45% cut in time to first token [1].

OpenAI also says it has reached its “AI research intern” goal. Its models now finish more than a third of day-long research tasks with no intervention, after failing most such tasks before [1].

Holding back a model. Altman confirmed OpenAI held back Astra 6.1 and paused training on a separate model to put more attention on safety, alignment, monitoring and security. Faster or cheaper versions of Astra can still ship [2]. He says OpenAI will go public eventually, but not while it is adjusting to new capability levels, and that investors are “very happy... and very patient” [2]. He wouldn’t comment on a reported \$70B ARR figure [2]. Bloomberg separately describes talks for a \$30B raise at a \$1.4T valuation [3].

Deals and funding

- **NVIDIA is acquiring Hugging Face.** Clem Delangue says the deal lets Hugging Face hire people it couldn’t as a small startup and “give them a decade to make open-source AI win” [8]. Together with AMD buying World Labs, both big chipmakers are now buying AI model and open-source platform companies.
- **General Intuition** raised another \$220M at a \$6.2B valuation as it starts releasing its models. It trains foundation models on billions of videos

labeled with actions, which it says can act in real time in environments they haven't seen [9].

- **Long Lake** completed its \$6.3B acquisition of Amex GBT, its 40th acquisition. Its companies now employ nearly 30,000 people, three years after it was founded [10]. It says its first cohort of companies doubled EBITDA in under two years while adding headcount [10]. This is the largest data point so far for the “buy a services business and deploy AI in it” model.
- **Listen Labs** is joining Salesforce AI Labs. Listen started as an AI interviewer and now recruits people, interviews them, analyzes the results and simulates how they'll behave. Customers include Microsoft, Anthropic and Sweetgreen. The deal came together while Listen was raising its next round [11].
- **Dorm Room Fund** closed a \$50M Fund 2 for student and recent-graduate or dropout founders. 30% of its investors are DRF alumni and their firms [12]. **a16z speedrun alpha** offers up to \$250K plus an automatic final interview for speedrun, which invests up to \$1M. The program runs in person January 11 to March 5 [13].

Policy and the big labs

David Sacks says frontier-lab leaders signed a **White House Accord on Super Intelligence**. By his account, they accept responsibility for safe development and take on new internal controls and external audits, which he presents as an alternative to an international agreement [14]. The post doesn't say what the audits involve.

Exponential View's read of **Anthropic's S-1**: an \$8B operating loss and a \$42B net loss in 2025, the latter inflated by an accounting charge. Revenue is growing much faster than costs [15]. It cites \$518B in compute commitments over 7–10 years, about \$410B of which cannot be cancelled, and expects annualized revenue above \$100B by the end of 2026. A November IPO is only a rumor for now [15].

Engineering signals

Harness design matters as much as model choice. Replit's agent lets the main model decide which specialist to hand work to, at what size, and with how much effort [16]. On DeepSWE it scores 72% at \$2.11 per task; Astra on its own scores 74% but costs \$4.43. Replit says neither baseline beats its agent on both cost and score. These are Replit's own benchmarks [16].

Shopify is moving its mobile apps from React Native to native Swift and Kotlin. Its reasoning is that agents now handle enough of the implementation, translation, testing and review that building for two platforms is no longer the deciding cost [17]. A shared test suite checks that the business logic behaves the same in both languages [17]. The implication is that tools built to

save engineering time by sharing code across platforms may lose some of their appeal.

Consumer and commerce

A Reddit summary of Menlo Ventures’ 2026 consumer report says global consumer AI spending rose from \$12B to \$40B while users grew only from 1.8B to 2B. In the US, the 14% of payers who spend \$100 or more a month bring in 60% of the money [18]. 32% of users have let AI act without their final approval [18].

a16z’s commerce essay separates two cases: agents that bring new customers, and agents that intercept orders a platform would have won anyway. Instacart is integrating with Muse, Amazon is trying to block it, and Shopify, Toast and Square gain because they have no marketplace ad business to protect [19].

Abliteration.ai says its GLM-5.3 release with customer-controlled guardrails drew customers from startups to Fortune 500 companies. It says they cited the big labs’ centralized guardrails as blocking their work, even inside special “cyber” programs. That is the company’s own claim [20].

Sources

1. OpenAI DevDay 2026 Keynote (FULL)
2. Altman on OpenAI’s IPO Plans, AI Agent Dots, AI Security
3. Bloomberg Businessweek Daily: OpenAI Dev Day (Podcast)
4. X post by @alliekmiller
5. X post by @thsottiaux
6. X article by @baseten
7. X post by @runwayml
8. X post by @ClementDelangue
9. X post by @gen_intuition
10. X post by @alextaubman
11. X post by @itsalfredw
12. X post by @themollette
13. X post by @andrewchen
14. X post by @DavidSacks
15. The first existential IPO
16. X article by @pirroh
17. Why has Shopify dropped React Native?
18. r/artificial post by u/AdBoth2756
19. X article by @aleximm
20. X post by @abliteration_ai