

OpenAI's \$70B run-rate figure turns out to be a gross-vs-net question, as Western open models pitch sovereignty

VC Tech Radar

2026-10-10

OpenAI's \$70B run-rate figure turns out to be a gross-vs-net question, as Western open models pitch sovereignty

By VC Tech Radar • October 10, 2026

The FT's ARR report on OpenAI shows how differently the frontier labs count revenue. Also: Reflection's Beam and the shift in token share toward open models, a16z's TypeSafe Series A, the cost of running agents, and constraints on infrastructure capital.

OpenAI's \$70B figure depends on how revenue is counted

Yesterday's brief reported OpenAI's run rate as approaching \$70B. On Thursday the *Financial Times* reported that the ARR figure OpenAI gave investors in its current funding round was actually \$20B below that widely cited number, and AI stocks fell [1]. Newcomer says the gap mostly comes from how the two labs count revenue. OpenAI reports net revenue. Anthropic reports gross, which includes models sold through partners like Amazon and Microsoft before those partners take their share [1]. Bloomberg later reported that OpenAI expects to reach \$70B annualized revenue by year-end, apparently on a net basis. Investors and customers see the two labs as roughly neck-and-neck [1].

This matters for the expected Anthropic IPO and for any comparison between OpenAI and Anthropic. Without adjusting for the accounting difference, yesterday's chart showing Anthropic overtaking OpenAI is hard to interpret. Newcomer links the same lack of transparency to deals priced at more than one valuation, which it says are becoming normal. It warns that unclear valuation figures carry real risks for investors, employees and founders [1].

Western open models are being pitched on sovereignty

Three open models launched this week, each pitched as the choice for buyers worried about AI sovereignty: Reflection’s Beam (US), Mistral’s Large 4 (France) and Aleph Alpha’s Kolibri (Germany) [1]. Beam has 501B parameters with 23B active. Large 4 has 1T parameters with 49B active and a 1M-token context window. Neither had released its weights yet; both said weights would come later this month [2].

Demand for the sovereignty pitch looks limited so far. Investors told Newcomer that few portfolio companies care where their open models come from, except those handling sensitive trade or government data. Many will use a Chinese model if a US provider such as Baseten or Fireworks hosts it [1]. The Western models match Chinese models on some benchmarks but trail on overall capability. GLM 5.3, Kimi K3 and DeepSeek V4.1 are still the most popular models on Baseten and Fireworks [1]. One investor added that customizing open models can get expensive quickly [1].

Reflection CEO Misha Laskin gave several data points on No Priors: - **Token share:** on gateways such as OpenRouter and Vercel, the token mix has flipped in about six months from 70/30 closed/open to 70/30 open/closed. He expects most token demand to go to open models, while closed-model companies stay very valuable [3]. - **Catch-up cost:** reaching the frontier cost hundreds of millions of dollars about 18 months ago and single-digit billions about six months ago. He expects around \$10B next year, roughly a 4x compute increase per model generation [3]. - **Beam’s training:** pretraining used 6,000 GB300s. The reinforcement-learning run used more than 10,000 GB300s for four weeks, so more compute went to RL than to pretraining [3]. - **Efficiency claim:** he says Beam is 3-4x more efficient than models in its capability class, and about 10x more efficient than some larger models [3].

A Lightspeed guest estimated that open-weight models are about 6–12 months behind the frontier. On that view, they suit high-volume work like support tickets, while long-running coding agents stay on closed models [2].

a16z leads TypeSafe AI’s Series A

a16z said it is leading TypeSafe AI’s Series A. Its thesis is that AI should “make software more powerful, and not just automate the recreation of existing software” [4]. A team member’s launch post says the model “accidentally served trillions of tokens a day,” that “29.4% of the fortune 500 showed up,” and that the company raised “a really big series A” three weeks after launch. The post gives no definitions or round size [5]. Founder Diogo Almeida pushes back on the idea that AI will wipe out SaaS: “software is very cheap” but not necessarily “easy to replicate.” He calls SaaS “one of the largest winners of the whole AI game” [6].

Agent costs depend on caching and routing more than list prices

Clouded Judgement argues that the cache-read rate, not the per-token list price, drives what agents actually cost [7]. In a modeled 40-turn session with an 85% cache hit rate: - Fable comes in about 16% below Astra, even though their list prices are the same. - Grok 4.7 is the cheapest flagship on list price but costs more than Sol and Terra [7].

The author suggests that cached context could create switching costs as agents run longer. He acknowledges this may be far-fetched [7].

Two operator data points: - **Asana** says that keeping an agent's page history stable enough for caching cut one browser-agent workflow from at least \$36.21 to \$0.47 per run [8]. The study used only three or four runs per condition, and its baseline included capped runs [8]. - **LangChain** says that sending each Open SWE task to the cheapest model that still passes quality tests cut median cost per task by 64% [9].

Infrastructure: capital, permits and power

- **Lambda** is reportedly raising up to \$4B at a \$14.5B pre-money valuation, possibly its last private round before a planned 2027 IPO [10]. Its backlog grew from \$15B in June to \$50B in September, but about \$35B of that appears to come from one Anthropic contract. Backlog is not revenue, and contracts can be cancelled [10]. Lambda has also raised debt several times [10].
- **Amazon** says it will stop using NDAs in data-center negotiations with local governments [10]. Communities are already passing temporary moratoria [10].
- **Pat Gelsinger** (former Intel CEO) says AI can design a chip in three months, but fabrication, packaging and rack integration mean nine months before it can be used [11]. He expects more data-center projects to default because the power won't be there [11]. He named vertical GaN power conversion and 800V DC as areas to watch [11].
- **Cisco** says its AI infrastructure orders went from zero to \$9.3B in two years [12].

Hard tech

- **Atomic Machines** came out of six years in stealth with the Matter Compiler. The company describes it as a system that builds micro-machines from code with no per-product tooling [13]. Its founder is Jeff Holden, who was Amazon's tenth engineer and Uber's first CPO [13]. Its first product is a 150-amp relay for 800V DC data centers, which it says it is beginning to ship to early customers [13].

- **Anduril** plans to invest \$3.7B in a submarine-component shipyard at Sparrows Point. The Navy contract is worth up to \$2.9B, with payments tied to production results [13]. Production in Maryland is targeted for 2030 [13].
- **Biohub** expanded a \$1.8B virtual-cell data collaboration. Google DeepMind, Isomorphic and Meta are putting in \$300M, and the total includes compute and existing data [13].
- **Ultra** reportedly raised \$50M to rent warehouse robots for a monthly fee rather than sell them. It has raised \$62M in total and its robots have packed more than 500K orders [14].

Venture structure

True Ventures’ Puneet Agarwal describes a split market. Capital is concentrated in fewer firms and companies, and early stage has “been forgotten a little bit” [15]. Because AI makes products easy to copy, True looks for founder depth and asks what could be proprietary or compounding, such as a data advantage [15].

On liquidity, SaaStr notes that the median tech company takes 11.5 years to IPO, while a standard venture fund lasts 10. Only 4% of US VC-backed exits in H1 2025 were IPOs [16]. Anthropic would take about 5.5 years from founding if it prices this year [16].

Elizabeth Yin described tranching SAFE rounds, where a founder offers a first tranche (for example \$300K of a \$1M raise) on better investor terms [17]. If the tranche fills quickly, the valuation was probably set too low. If it fills slowly even at a low valuation, the raise likely has a bigger problem [18].

The UK plans measures to functionally ban non-competes for startups and scale-ups, with details due at the October 28 budget [19]. The announcement opens a consultation and does not settle the policy [19].

Science signals

- **OpenAI math results:** no one has reviewed all of OpenAI’s 700+ math manuscripts, not all are Lean-certified, and some proofs have already been retracted after outside scientists found errors [20].
- **Station:** in this agent environment, agents rediscovered 62.7% of the findings from three ICLR oral papers, compared with 14.4–20.6% for AI Scientist-v2 [21].
- **Senolytics:** a 31-patient phase 2 trial of senolytic drugs in fibrotic liver disease (MASH) reported fibrosis improvement in 47% of patients versus 7% on placebo [22].

Sources

1. Competing ARR Estimates & Fuzzy Valuation Math in the Spotlight as Markets Get Edgy
2. Reflection Beam, Mistral Le Chonk & the Drone That Delivers Your Chipotle | Lightwork
3. Beam: The Great American Open Model with ReflectionAI Co-Founder and CEO Misha Laskin
4. X post by @a16z
5. X post by @CompleteSkeptic
6. X post by @a16z
7. Clouded Judgement 10.9.26 - Cache Is the New Egress
8. r/artificial post by u/Codeblix_Ltd
9. X post by @hwchase17
10. Amazon and others are done keeping their data center deals secret | Equity Podcast
11. Former Intel CEO: Why This is the Best Time to Build Hardware
12. X post by @ycombinator
13. Weekly Dose of Optimism #214
14. X post by @ritwikpavan
15. Build Mode Bonus: Inside the VC Decision-Making Process with True Ventures' Puneet Agarwal
16. It Takes 12 Years to IPO. Then Another 2-3 for Your VCs to Get Liquid. And Founders Can Generally Only Sell About ~4% a Year.
17. X post by @dunkhippo33
18. X post by @dunkhippo33
19. X post by @Dom_Hallas
20. r/artificial comment by u/scientificamerican
21. r/MachineLearning post by u/progenitor414
22. r/Futurology post by u/lughnasadh