Grok Bot hits a cost wall as Big Tech capex nears $745B and lenders want a backstop

VC Tech Radar

2026-11-02

Grok Bot hits a cost wall as Big Tech capex nears $745B and lenders want a backstop

By VC Tech Radar • November 2, 2026

AI's unit economics showed strain from two directions this period: subsidized agent subscriptions started getting rationed, and Big Tech's buildout is increasingly paid for with debt and equity. Also covered: robotics after Astra, agents replacing vendors, and VC debates on tranched rounds and "pre-consensus" investing.

Subsidized agent pricing starts to break

Some Grok Bot users ran through a week's usage overnight [1]. Elon Musk's explanation was that "the cost of serving Grok Bot was unsustainably high relative to revenue" [2]. Amjad Masad drew the wider lesson: "If you're enjoying a subsidized AI subscription get ready to get rug pulled" [3]. OpenAI is also reportedly capping how many paid ChatGPT accounts one person can hold per device and payment card [4].

Jason Calacanis read the same signal as a sign of demand. He called it "an extremely bullish sign that people are getting value from their bots" and valued Grok bots at about $30 an hour of employee-equivalent work [5]. He says he spends under $40k a year on AI tools, gets more than $2M in value, and has automated "at least seven full-time jobs" [6]. Both readings can hold at once: agents are producing real value, and the companies serving them are not covering their costs. For portfolios, the risk is gross margins built on flat-rate plans with heavy agent use. The prior brief's theme of cheaper tokens but larger bills now shows up as rationing by providers.

The capex math, and who is financing it

Rob May at Investing in AI says Amazon, Alphabet, Microsoft and Meta now guide to about $720–745B of 2026 capex, up from about $410B in 2025 [7]. In Q2, cash capex was 115% of operating cash flow at Alphabet and 119% at Amazon, and long-term debt at both has roughly doubled since December 2025 [7]. He sets Anthropic's $16.8B Q3 revenue and OpenAI's roughly $50B run rate against that spending. His conclusion: "Revenue is real and the gap is real" [7]. He also cites Epoch AI data showing the price of reaching a fixed level of model performance falling about 50x a year. In his view, value is shifting to whoever owns the scarce inputs and the customer relationship [7]. His screen for investments: a company should own at least two of three things — a binding constraint, proprietary domain data, or a workflow customers can't easily leave [7]. He discloses that his own routing business benefits if this thesis is right [7].

SaaStr's notes on the week's 20VC episode add more financing detail: - Banks and Blackstone are providing $63B to buy Broadcom chips for Anthropic, and Broadcom is guaranteeing part of the chips' resale value to senior lenders. SaaStr called this the week's most worrying number [8]. - Meta launched about $35B of equity offerings, backed by a $60B at-the-market program, after its first negative free cash flow in more than a decade [8]. - Anthropic disclosed roughly 900% net revenue retention (NRR). SaaStr estimates the durable figure is closer to 300–400% once pilots converting to production are excluded [8]. - Rory's reading of the S-1 is that about 80% of Anthropic's revenue comes from models less than six months old [8].

The advice to founders is not to count on venture debt or compute financing at good terms in 2027 plans [8]. On supply, Bindu Reddy expects spot GPU prices to reach 3–5x existing contract prices. If so, neoclouds locked into low-priced contracts would lose out [9]. A Reddit post relays a SemiAnalysis headline warning that 2028 data-center buildouts face serious risk of pausing after the US midterms. No details beyond the headline were visible [10]. Hunter Walk puts a "nonzero chance" on the US government becoming an equity or debt buyer of last resort for Anthropic, OpenAI or neoclouds if the AI market wobbles. He cites the Intel golden share as precedent [11].

On the hard-tech side, Garry Tan called YC "the YC for hard tech" [12]. He was pointing to Viraj Aero (YC F26), which adapts a flight turbine to power data centers. The company says the turbine makes its own cooling water from exhaust, gets 40% more power from the same fuel, and can be permitted in weeks [13].

Altman in Abu Dhabi: an unverified Millennium Prize claim

Speaking at an ADNOC event, Sam Altman said 10,000 agents of the "next GPT" worked for about four days with no hierarchy "when we proved the first

Millennium Prize problem" [14]. He did not name the problem or say how the proof was checked. Given the disputed math claims in the prior brief, treat this as unverified. In the interview that followed, he said global average use is now about 100,000 tokens per person per month. He called energy "the very dominant factor" in AI's future cost [15]. The host described ADNOC as the first energy company to adopt OpenAI Frontier, as part of agreements with G42 [15].

AI-found vulnerabilities keep piling up

Martin Casado pointed to a remote-code-execution flaw found in the Linux TCP stack. Searching kernel code used to not be worth the effort, he said, "But with AI, it's basically free" [16]. Altman argues that AI is exposing vulnerabilities that already existed and that security-fix practices need to change [17]. The linked article reports that Oracle's latest quarterly patch addressed about 3,000 issues, most of them found with AI. Oracle is now partnering with OpenAI on vulnerability discovery [18]. That points to a market for automated patching and attack-surface inventory.

Robotics: is Astra the "GPT-3 moment"?

Freda Duan, summarizing talks with 50 people in robotics, argues that Astra is robotics' GPT-3 moment and that robotics is becoming "just another vertical for frontier labs" [19]. Garry Tan amplified it: "Practical robotics is nigh" [20]. Her reported signals: - OpenAI reorganized its 30–40-person robotics team around Astra. - There is talk of data-procurement budgets approaching $1B per frontier lab [19].

She expects the moat to shift to reliability, integration and deployment [19]. That puts pressure on robotics foundation-model startups, whose pitch she calls a "narrower value proposition" [19]. She also says Astra is about a month old and it is too early for firm conclusions [19].

Agents as buyers: churn risk for B2B software

A core SaaStr vendor went down for a full day. SaaStr's AI "VP of Marketing" agent, 10K, set up a replacement and recommended never using the vendor again. SaaStr switched at least partly that same day [21]. 10K had earlier replaced Notion, and a heavily throttled API drove SaaStr off Marketo [21]. SaaStr's takeaway for vendors: "Your API is the product," and sunk-cost prepayments protect much less when customers' agents can switch in hours [21].

DeepSeek's Harness team open-sourced an MIT-licensed "super-agent" prototype. A persistent Router assigns tasks to isolated Worker sessions, and by default the system only observes on work it infers rather than acting [22]. Its reported speedup came from a simulation with synthetic messages and did not measure code quality [22].

Venture: tranched rounds and "pre-consensus"

David Tisch, speaking on 20VC, warned that VCs who structure tranched rounds can push co-investors "to a different product than you get to invest in." He added that he has "a long memory" [23]. He also argued tranches can help some founders through more runway, momentum and buzz for recruiting, and called the lean startup "a total fake" [24].

Hunter Walk argues that truly non-consensus seed deals often can't raise follow-on rounds from major VCs. His advice to emerging managers is to be "pre-consensus": investments should become consensus within 6–24 months [25]. Paul Graham warned that investors underrate how much simply talking to them distracts a startup, even for YC companies [26].

Other signals

- **Courts:** Shenzhen says its judges each closed 744 cases in 2025 with an AI assistant that drafts rulings, 249 more than in 2024. Sir Geoffrey Vos, who just retired as head of England and Wales' civil courts, said simple cases "can and will be decided by AI if the parties agree" [27].
- **Commerce:** a16z says shoppers referred by AI convert 60% better and bring 53% more revenue per visit than other online shoppers [28].
- **Math talent:** an MIT survey reportedly found nearly 76% of math grad students and postdocs are increasingly unlikely to stay in academic math [29].
- **Cheaper models:** Garry Tan claims Particle-1 matches much larger models on agentic retrieval at 1/23 the price [30].
- **Deadline:** YC W27 applications close Nov 2 [31].

Sources

1. r/artificial post by u/ross2000
2. X post by @elonmusk
3. X post by @amasad
4. r/artificial post by u/Alone-Dragonfruit602
5. X post by @Jason
6. X post by @Jason
7. Silicon Valley Is Wrong About AI, And Wall Street Is Right
8. Jason's Takes on This Week's 20VC: Anthropic's S-1, Microsoft's 27% Buyout, and My 100-Day Rule
9. X post by @bindureddy
10. r/artificial post by u/Eliv_nurotic
11. Will the US Government Become AI's (SI?) Lender of Last Resort?
12. X post by @garrytan
13. X post by @joseph_risson

14. FULL SPEECH: OpenAI CEO Sam Altman Reveals How AI Could Transform the Future of Global Energy | AF14
15. FULL INTERVIEW: OpenAI CEO Sam Altman Reveals What Excites Him Most About the Future of AI | AC14
16. X post by @martin_casado
17. r/Futurology post by u/Gari_305
18. r/Futurology comment by u/Gari_305
19. X post by @FredaDuan
20. X post by @garrytan
21. Our Core Vendor Was Down for 24 Hours. Our AI Agent Replaced Them Before They Came Back Up.
22. r/artificial post by u/winer666
23. X post by @HarryStebbings
24. X post by @HarryStebbings
25. For Early Stage VCs, Non-consensus is Too Risky, Consensus is Too Boring, but Pre-consensus May Be Just Right.
26. X post by @paulg
27. r/Futurology post by u/mawcopolow
28. X post by @a16z
29. r/Futurology comment by u/Gari_305
30. X post by @garrytan
31. X post by @ycombinator