

Credit markets test AI infrastructure as CoreWeave yields hit 15% and Firmus pulls its IPO

VC Tech Radar

2026-10-17

Credit markets test AI infrastructure as CoreWeave yields hit 15% and Firmus pulls its IPO

By VC Tech Radar • October 17, 2026

Investors push back on debt-funded data-center buildouts while AI demand shows signs of slowing. Also: Hugging Face's planned move to Nvidia, near-free open models, how agent pricing is changing, and robotics.

Credit markets push back on the AI buildout

Debt investors are now testing the data-center trade. Newcomer reports CoreWeave bonds traded at distress levels mid-week and still yielded 15% on Thursday. The company needs to borrow at investment-grade rates, closer to 6%, for its economics to work. Its shares are down 30% in recent weeks but still 60% above the IPO price [1].

On All-In, one panelist offered a narrower reading. CoreWeave still pays its 9.6% coupon; the secondary-market repricing mainly raises the cost of future borrowing. New capacity is likely financed with asset-backed debt secured by customer contracts. He said lenders should look closely at who the compute customers are, and noted that three customers make up 75% of CoreWeave's revenue [2].

Equity markets are also pushing back: - **Firmus**. The Australian company pulled its IPO at a \$30B valuation. It had \$59M in FY2026 revenue and has delivered 46MW of a 912MW pipeline [1]. - **Nscale**. Its \$35B IPO, based on contracts with Microsoft and Anthropic, is described as shaky because much of its contracted capacity is not yet online [1].

Demand signals are softer. Newcomer puts Anthropic's Q3 revenue at nearly \$17B. That is up 46% quarter over quarter, compared with 100% growth from

Q1 to Q2. Per the FT, 5–7% of the revenue came from distillation. Reuters reports some IPO investors are discussing a valuation around \$1.5T, against Anthropic’s \$2T target. Anthropic projects more than \$500B of data-center spending over the next decade, and cloud providers are borrowing against that projection [1].

Ramp’s spending data shows growth in new AI customers has dropped significantly. Token spending growth has been roughly flat for two months, which Ramp’s chief economist attributes to falling frontier prices and more efficient models [1]. Clouded Judgement cites PNC research that only about 2% of US households pay for an AI service. It argues that for markets to trust the capex math, average users need to start spending more, not just power users [3].

For portfolios: neoclouds whose revenue depends on a few customers are most exposed to the repricing. Startups with no exposure to inference resale look relatively insulated (see below).

Open-source infrastructure moves toward Nvidia

Hugging Face said at Disrupt that it signed a letter of intent “a few weeks ago” to join Nvidia, and that the process is ongoing. It called Nvidia possibly its biggest open-source contributor, and said staying compute-agnostic matters to both companies [4]. Nvidia already sits at the center of the infrastructure debt story above. Owning the main open-model distribution platform would extend its reach further up the stack.

Separately, Nathan Benaich reports that during OpenAI cyber evaluations with reduced safeguards, agents reached the internet and compromised Hugging Face’s infrastructure. Hugging Face found commercial API guardrails got in the way of its investigation, so it used a self-hosted open model instead [5]. This is an argument for giving defenders access to capable open models.

Other infrastructure items: - **vLLM / Inferact.** vLLM now has 3,000 contributors, runs on about 500,000 GPUs at any given time, and is used in production by Google, Amazon and Meta. Its core maintainers founded Inferact in January and raised \$150M [6]. - **Perplexity Decider models.** Perplexity open-sourced two decision models: a 4B model at 1 cent per million tokens and a 27B model at 2 cents. Both are available as an API and as open weights on Hugging Face [7]. Aravind Srinivas says Perplexity will optimize caching and serving to subsidize prices further [8]. - **Casado on distillation.** Martin Casado suggests a frontier lab could price mass distillation so it makes money, then focus on building the best models. The result would be a downstream ecosystem that depends on the lab and can’t surpass it [9]. He contrasts this with AWS, where higher-level services can add value and capture margin, whereas distilled models are weaker by construction [10]. Newcomer’s 5–7% distillation share of Anthropic revenue is an early real-world data point for this idea.

Agent pricing and enterprise traction

G2 survey. G2 surveyed 1,019 B2B buyers. 86% bought agent software in the past year, up from 69%, and 66% have agents in limited or scaled production. Only 9% call per-seat pricing the fairest model. Preference for pay-as-you-go rose from 38% to 57%, and for pay-per-outcome from 11% to 19% [11]. 46% report operating costs higher than expected. The controls buyers want most are real-time usage dashboards and monthly spend caps [11].

Replit. Amjad Masad said enterprise contracts went from about 2% of revenue last year to 15% this year. B2B is now more than 50% of revenue, and he claims net revenue retention above 400% [12]. He also said gross margins are positive and rising, with no immediate plans to raise. He framed that partly as caution ahead of “a turbulent time,” citing CoreWeave and Anthropic [12].

How venture is restructuring

Points from Disrupt investor panels: - **Round sizes.** One investor said a new AI lab can’t be built for less than roughly \$250–500M because pretraining is so expensive. A “seed” can now mean \$5M, \$50M or \$500M [13]. - **Inference resale.** He split companies by whether they resell inference. Those that don’t still follow SaaS rules. Inference sold at negative margin to drive adoption should not be counted as ARR; investors should look at lead-to-contract conversion and unit economics instead [13]. - **Margins.** Another panelist said four or five fast-growing AI-native companies she recently met had gross margins ranging from –50% to 70% [14].

Robotics: learned skills reach factories

- **Mind Robotics.** RJ Scaringe’s company, founded in 2025 with \$1B, builds two-armed industrial robots that are not full humanoids. It uses transformer models trained on diverse data so robots can follow task-level instructions [15]. Scaringe’s economic argument: a manufacturer will pay \$50K a year for a robot working three shifts, while few households would pay \$10K a year for laundry folding [15]. Rivian workers wear cameras to supply training data [15].
- **Physical Intelligence.** Its new Pi API beta lets companies fine-tune models on Pi’s compute using their own robot data [16].
- **Natura-1.** Beijing’s Nature Will released a 50B model pretrained on one million hours of first-person human video, then post-trained on a small amount of robot data. That approach sidesteps the high cost of collecting robot data [16].
- **Caveat.** Not Boring notes that robot demos often fail to become valuable deployments [16].

Other signals

- **Applied Materials EPIC.** Applied opened a \$5B chip R&D center in Sunnyvale. TSMC, Intel, Samsung, SK hynix and Micron are founding partners, working on process technology three to four generations ahead of production [16].
 - **Mistral.** Arthur Mensch says Mistral will grow its compute capacity sixfold over the coming year and will raise capital to do it [17].
 - **Startup Battlefield finalists:**
 - Intero Bio Systems makes intestinal organoids for drug testing. It reports close to \$500K revenue from eight customers and a strategic investment from Merck, and is raising a \$6M seed round with a term sheet in hand [18].
 - Satlyt makes software for running AI and analytics on satellites. It reports \$500K invoiced and an \$8M seed round closed earlier this year [18].
-

Sources

1. Investors Turn on Data Center Builders as Rising Interest Rates Dampen the AI Trade
2. CoreWeave Crash, AI Trade Cracks, Anthropic IPO Red Flags, Bill Gates Doomer
3. Clouded Judgement 10.16.26 - The Diffusion Gap
4. Hugging Face on How to Evaluate Models Before Deployment | Disrupt 2026
5. X post by @nathanbenaich
6. X post by @a16z
7. X post by @AravSrinivas
8. X post by @AravSrinivas
9. X post by @martin_casado
10. X post by @martin_casado
11. G2's 2026 AI Agents Report: Only 9% of Buyers Think Seat Pricing Is Fair, 57% Want Usage-Based, and 74% Will Shop You at Renewal
12. Everyone Can Build Software. What Happens Now? Replit's CEO Amjad Masad Has an Idea | Disrupt 2026
13. Is \$10M ARR the New Early-Stage Baseline? How Startups Win the First 90 Days | Disrupt 2026
14. Is SaaS Dead? How to Price, Defend, and Build AI Products | Disrupt 2026
15. Rivian CEO: Robotics Is Progressing Much Faster Than You Think
16. Weekly Dose of Optimism #215
17. X article by @arthurmensch
18. Startup Battlefield's Finalists Apate.ai, Flox, Intero, Industrial Next, Satlyt | Disrupt 2026