

Agents Are Becoming Both the Customer and the Competitor

VC Tech Radar

2026-08-18

Agents Are Becoming Both the Customer and the Competitor

By VC Tech Radar • August 18, 2026

The strongest signals are strategic capital moving into physical AI, agents beginning to replace narrow B2B products, and a widening split between cheap specialized intelligence and expensive frontier-model economics.

1. Funding & Deals

Uber’s investment in Zipline is a distribution-led physical-AI bet, not a clean early-stage comp. Zipline says Uber is becoming an investor while it scales to more than 1 million autonomous Uber Eats deliveries per day. [1] Jason Calacanis says the partnership gives Zipline access to Uber Eats’ existing partner network, customer support, and customer-acquisition infrastructure; he also discloses that he put millions into a recent late-stage Zipline round, so the current evidence supports a strategic-capital signal rather than a Seed/Series A valuation benchmark. [2]

The diligence question is whether the urban last-mile model travels beyond Zipline’s rural medical-delivery track record: public discussion explicitly distinguishes those environments, while another response flags noise and privacy as adoption risks. [3, 4]

2. Emerging Teams

TryNearbyCom has the clearest current early-stage traction signal. The YC S26 startup says it is live with more than 120 paying restaurants across Southern California; after 10 months of restaurant visits with creators and conversations with hundreds of owners, it reports more than 130% growth since the batch started and over 90% retention since November. [5] The signal is a

potentially repeatable local-creator distribution model; before underwriting it, verify the retention cohort, revenue quality, and restaurant-level payback.

Forge is a founder-signal rather than a traction story. A 17-year-old developer says he has built a C++ deep-learning framework from scratch since January, loaded real GPT-2 weights into a Forge implementation, and matched Hugging Face’s output token-for-token. [6] The project is still CPU-only, lacks a KV cache, and is working toward CUDA and performance fixes, so the investment question is whether the unusual systems depth can become a team and product rather than whether this is already a company. [6]

Agent Facets is an early agent-supply-chain thesis. Its builder argues that public agent “skills” should be managed as dependencies, with version ranges or pins, immutable artifacts, integrity checks, and reproducible installs across Claude Code, Codex, OpenCode, and other clients. [7] The motivation is credible infrastructure pain: the post cites reports of malicious skills and a Snyk figure that 36% of scanned skills contained prompt injection, while explicitly saying that figure may not be fully accurate. [7] The current signal is the category definition—portable, reviewable agent capabilities—not disclosed revenue.

3. AI & Tech Breakthroughs

A JAMA study puts autonomous clinical AI on the agenda, but not yet in production. In a public summary of the paper, Khosla and coauthors report 159 simulated OSCE cases in which physicians rated Google’s AMIE better than physicians at eliciting complaints (97% vs. 50%), systems review (88% vs. 35%), medical history (85% vs. 50%), family history (50% vs. 21%), and medication history (68% vs. 45%). [8] The authors argue that AI-alone may eventually outperform physician-only or physician-AI hybrids in some cognitive workflows, while listing workflow, liability, regulation, reimbursement, and medical education as unresolved barriers and pointing to possible deployment in some workflows by 2030. [9] Because the evaluation is simulated and the claims here come through the authors’ summary, the near-term investment signal is in evaluation, governance, and clinical deployment infrastructure—not proof that autonomous care is ready.

Faraday is a more concrete step toward scientific agents, while open-ended discovery remains a separate gate. The paper’s abstract describes Replica, a scalable paper-replication task space with an auto-generated rubric judge, and Faraday, a 27B agent that uses coding agents as tools and surpasses Claude Opus 4.8 and GPT-5.5 on held-out replication tasks. [10] Import AI reports that Replica covers 100 ML and AI-for-science papers converted into 310 tasks, with Faraday exceeding the comparison systems on 73% of in-distribution ML tasks and 60% of held-out AI-for-science tasks according to its rubric judge. [11] By contrast, DiG-bench’s 70 hidden-rule games remain difficult for frontier models: only Opus 5 and Fable 5 with Claude Code beat any Tier 7 task, while

individual humans reached 100% on the tests. [11] Replication may therefore be an earlier commercial wedge than genuinely open-ended scientific discovery.

CR-NN is an unverified low-memory attention bet. Its author claims matrix-free attention with $O(N \log N)$ complexity and a $16.2\times$ speedup over Flash at 50K tokens, plus 0.015 GB at 1.36M tokens versus 12.3 GB for a KV cache; the project includes negative results and is explicitly seeking collaborators to validate the idea at scale. [12] Treat the numbers as a replication target, not an established efficiency result.

4. Market Signals

AI demand is expanding, but usage is highly concentrated and frontier-model spend may be nearing a ceiling. Exponential View reports July AI revenues at three times the prior year’s level and an annualized run-rate above \$210 billion. [13] Its tracking also says the top 10% of OpenAI enterprise users consume $8.3\times$ as many tokens as the typical firm, while Fable 5 usage is flat at only 6% of business tokens and 11% of spend—an attributed signal that the best model is not automatically the economic default. [13]

Agentic replacement is now visible in churn, while some incumbents are using agents to expand. SaaStr says it canceled Notion after seven years because its internally built 10K agent absorbed Notion’s remaining job, running the Monday staff meeting from revenue, campaign, collections, and pipeline data. [14] It argues that quiet, low-touch accounts are especially exposed because a narrow workflow is easier for an agent to absorb, and says Marketo lost a 10-year relationship after its API stopped working for SaaStr’s agents. [14] The counterexample is Stripe: it says agents wrote 30% of its code in a week, cut global tax-filing time to one-third of the U.S. version, made sellers 20% more productive, and led it to hire more sellers. [15] The practical split is between AI that unlocks new capacity and AI that makes a narrow incumbent product unnecessary.

Open-model economics are becoming a strategic fault line. Interconnects describes open-model training as highly capital intensive and says Nvidia is reportedly spending \$26 billion to create a broad ecosystem of model builders and inference demand, while acknowledging that it is unclear whether the strategy will pay off. [16] Its base case is a bifurcation: closed labs retain the most valuable knowledge-work, drug-discovery, and software-engineering markets, while open models specialize in efficient, modifiable, enterprise-specific agents running on private data; revenue-share licenses are being tested to keep near-frontier open-weight development financeable. [16]

Agent commerce will probably reuse existing payment rails, leaving authorization and exception handling as the wedge. A SaaS discussion favors agents using ordinary checkout rather than requiring every merchant to build a new API, but identifies CAPTCHA/3DS flows, subscription permissions, and fraud or chargeback liability as unresolved problems. [17, 18, 19, 20, 21]

Perplexity’s current product direction points to the same control layer: users can set connector tools to Allow, Always Ask, or Deny, with recurring runs following thread-level approvals; CEO Aravind Srinivas frames this as keeping humans able to intervene. [22, 23]

5. Worth Your Time

- **Watch — Will Gaybrick on a16z.** The conversation connects Stripe’s “build everything” posture, its 7,000 one-shot PRs per week, disappearing checkout pages, and stablecoin-enabled micropayments. [15, 24]
- **Read — Teaching Everyone to Fish for Tokens.** The clearest current framing of Nvidia’s open-model strategy, the financing problem for open-weight labs, and the likely shift toward specialized on-prem agents. [16]
- **Read — Training AI Scientists to Replicate Research.** Start with the original abstract for Replica’s rubric-based evaluation and Faraday’s held-out replication result before accepting the broader AI-scientist thesis. [10]
- **Try — GBrain.** Garry Tan’s free, MIT-licensed project generates a personalized agent for Codex or Claude Code through a 12-question onboarding, installs 70 skills, and creates a private knowledge wiki. [25, 26, 27]

Sources

1. X post by @Keller
2. X post by @Jason
3. r/Futurology comment by u/frank_datank__
4. r/Futurology comment by u/1776FreeAmerica
5. X post by @FaZeApex
6. r/deeplearning post by u/Express-Act3158
7. r/SideProject post by u/TheCritFisher
8. X post by @vkhosla
9. X post by @vkhosla
10. Training AI Scientists to Replicate Research
11. Import AI 469: Science AI; RSI simulator; and Zuck’s technological pessimism
12. r/deeplearning post by u/edisonwine
13. Data to start your week
14. We Churned Notion After 7 Years. Our AI Agent Took Its Last Job.
15. X post by @a16z
16. Teaching Everyone to Fish for Tokens
17. r/SaaS post by u/No-Echo-6522
18. r/SaaS comment by u/CheerfulTycoon

19. r/SaaS comment by u/Feisty-Sock-2565
20. r/SaaS comment by u/Able-Mud5277
21. r/SaaS comment by u/Few-Ocelot-3271
22. X post by @AskPerplexity
23. X post by @AravSrinivas
24. X post by @a16z
25. X post by @garrytan
26. X post by @garrytan
27. X post by @garrytan