

AI's Next Investable Layer Is Usage, Control, and Workflow Ownership

VC Tech Radar

2026-09-07

AI's Next Investable Layer Is Usage, Control, and Workflow Ownership

By VC Tech Radar • September 7, 2026

The most actionable signals this period sit above the base model: consumption-linked economics, agent governance, exact inference layers, and vertical workflows with measurable ROI.

1. Funding & Deals

The financing signal this period is a rotation toward usage-linked revenue and security exposure. A SaaS market review reports strong performance from CrowdStrike, Twilio, Snowflake, and Datadog, arguing that each either bills by consumption or sells into budgets expanded by AI; it also notes the downside of concentration when a large customer cuts usage. [1] The same review contrasts that with declines in seat-priced HubSpot, monday.com, and Figma; Figma's AI credits increased inference costs while its pricing remained per editor. Its conclusion is that the billing unit—not growth alone—did much of the sorting as CIOs reallocate budgets toward tokens. [1] For early-stage deals, the diligence question is whether revenue and gross margin scale with customer usage or remain exposed to seat-budget optimization. [1]

Seed robotics capital is now confronting an operating bottleneck. An unnamed seed-stage robotics startup reports \$4.2 million in funding and a technically dense team of three Robotics PhDs, a BME PhD, a MechE PhD, two robotics MS holders, and a BME MS holder; its advisers include technical executives associated with billions in exits. The CEO is also acting as CTO, and the company says it needs a systems engineer or technology lead to set priorities and review acceptance criteria. A senior BME systems engineer reportedly rejected an offer of \$130,000 plus 0.5% equity as too low. [2] The investment issue is executional: technical pedigree is not yet translating into a repeatable decision

and quality-control layer that can scale beyond the founder. [2]

2. Emerging Teams

A vertical AI operating tool is starting from measurable margin leakage rather than a generic chatbot. A 40-year-old family commercial-cleaning business with roughly 40 accounts connected payroll hours, billing, expenses, wages, overhead, and pricing data, then found that some long-running accounts were barely profitable or loss-making. The founder reports that money retained after costs rose from roughly \$30,000–\$40,000 on \$3 million of annual revenue to roughly \$80,000–\$90,000 without doubling the customer base. The proposed product would use each operator’s own financial history to flag weak accounts, recommend low/base/high-margin bids, and compare estimates with actual hours and costs. It is still a validation project, not a launched company. [3] A few owners have already reached out; the decisive test is whether another operator can obtain useful results without the founder explaining the system personally. [4]

A niche data/IQ platform has strong repeat usage, but its buyer is not the crowd. The founder reports 150 forecasting votes in seven days, with 130 repeat voters and 60 voting three times; separately, a B2B user said they would pay for three quarterly opportunities unavailable through LinkedIn Sales Navigator or Crunchbase, while a Big Four user has used the due-diligence feature twice monthly for three months. [5] The more actionable signal is the specific paid outcome and recurring professional workflow: feedback warns that forecasting, opportunity sourcing, and due diligence may be three different products, and that the recurring due-diligence user is a better monetization test than the vote count. [6, 7]

3. AI & Tech Breakthroughs

Inference efficiency is moving into the runtime layer. A prototype Sliding Window Attention layer applies attention sinks plus a recent-token window to pretrained Hugging Face models without retraining. In a Qwen2.5-7B experiment, reported KV-cache memory stayed near 3.5 MB at both 16K and 32K context, versus about 923 MB and 1.84 GB for full attention; full attention ran out of memory at 64K while the bounded version remained usable. TPOT fell from roughly 38.4 ms to 30.5 ms at 16K in that setup. The trade-off is material: tasks needing information outside the active window can degrade, and the author is still separating inherent limitations from implementation or model effects. [8]

Uncertainty estimation may become cheap enough for serving, but agreement can still be confidently wrong. A released benchmark reports that normalized exact-match entropy reached 0.889 AUROC on GSM8K across 7B–27B models, matching neural semantic entropy while running in under 2 ms on CPU; the neural approach reportedly takes 100-plus seconds on CPU

and adds substantial VRAM overhead. On 120B models, however, the authors describe “Confident Mode Collapse”: identical incorrect answers across samples drove AUROC down to 0.091. These are benchmark results, not evidence of a solved reliability problem. [9]

THREADS is a useful hybrid-system pattern: exact memory below a neural model. The open-source prototype accepts structured facts and relationships, tracks temporal changes, retractions, contradictions, and provenance, and reports exact results on tests including a 200,000-hop chain, a 128-hop query amid 1 million irrelevant events, 5,000/5,000 historical queries, and 40,000/40,000 ambiguity/contradiction cases. Its author explicitly says it is not a replacement for transformers, databases, or SMT solvers and does not understand arbitrary English; the proposed role is an exact memory/reasoning layer beneath a language model. [10]

4. Market Signals

The security-response window is collapsing at the same time that agent governance is lagging. a16z reports that roughly 87% of software bugs hackers exploit are attacked on or before the day the bug becomes public, up from 23% in 2020. [11] Separately, a practitioner post argues that enterprise security teams are shipping their own threat-detection and triage agents faster than governance teams can track; it claims shadow agents are already common, with tool access, data connections, and runtime behavior rarely monitored centrally, while internally built agents may sit outside the 80-plus regulatory and security frameworks enterprises use. [12] The investable control-plane opportunity is therefore not just better model detection, but faster vulnerability intelligence, agent observability, and enforceable permissions.

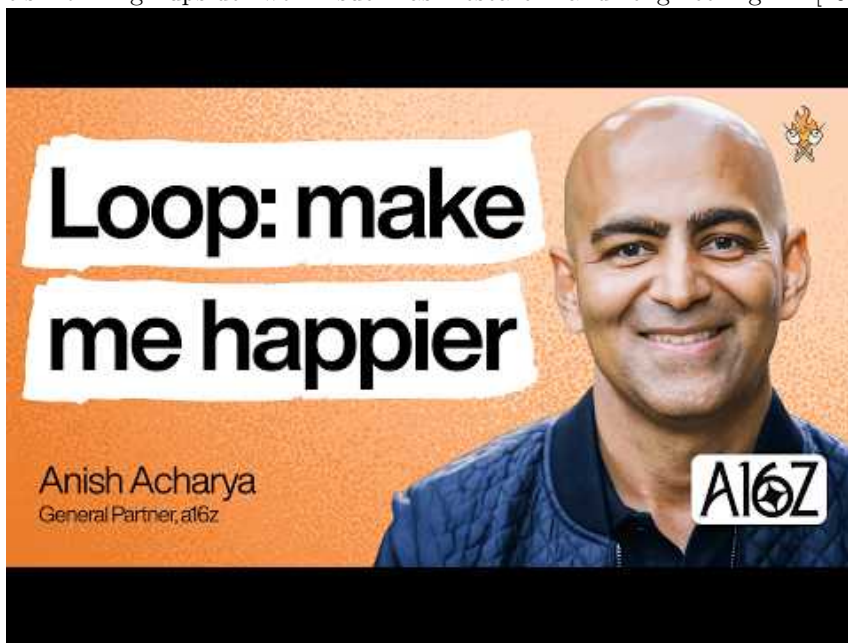
AI infrastructure is creating real labor demand while increasing land and permitting risk. An a16z post relaying Goldman attributes more than 300,000 construction jobs to the AI buildout since 2022, including roughly 75,000 in the past year; electrician and HVAC trades are growing about 2% annually, roughly twice the rate of construction overall. [13] At the same time, a current post says rural land prices are rising as some owners sell into data-center development while others resist it. At a July protest in Lubbock, Texas, Agriculture Commissioner Sid Miller was quoted saying developers were taking prime farmland and sometimes offering up to 10 times its value. [14, 15] Land acquisition, grid access, permitting, and local support now belong in the deployment case alongside power and chips.

Coding capability and coding economics are sending different signals. Martin Casado says recent models show a meaningful step in computer use but no comparable improvement in coding for his work, and suspects further model advances for high-skilled development tasks may no longer make economic sense even as cybersecurity improves. [16, 17] A separate builder reports that the most capable coding models sit behind expensive tiers, consume more

tokens, and can cost more than \$100 to improve a single function, potentially giving venture-backed and corporate teams an advantage over budget-constrained developers. [18] The cost report is anecdotal, but it supports a narrower underwriting question: where does agentic coding create enough accepted output to justify frontier-model spend?

5. Worth Your Time

- **Watch — Why companies are becoming a series of loops | Anish Acharya.** The useful segment treats agents as loops across coding, growth, sales, support, and legal, but says loops plateau at local maxima and still require human out-of-distribution judgment. The same conversation offers a practical model mix: cheaper open-weight models for bounded tasks and expensive frontier models for high-upside work such as research and engineering. [19]



Why companies are becoming a series of loops | Anish Acharya (a16z) (11:50)

- **Read — The “Owning” Phase of AI, Part 1.** The essay is a useful screening framework for vertical-model companies: proprietary non-scrappable data, an API-cost tipping point, low tolerance for mistakes, workflow ownership, and privacy mandates. It cautions that a custom-model announcement is not enough; the investable combination is captive data, captive distribution, and sufficient scale. [20]

Sources

1. Beyond the SaaSocalypse: The Winners and Losers of 2026 (So Far)
2. r/startups post by u/climbingTaco
3. r/EntrepreneurRideAlong post by u/Equivalent-Catch9111
4. r/EntrepreneurRideAlong comment by u/Equivalent-Catch9111
5. r/SaaS post by u/Candle_Realistic
6. r/SaaS comment by u/West_Inevitable_2281
7. r/SaaS comment by u/MiserableDocument509
8. r/MachineLearning post by u/ahsaor8
9. r/deeplearning post by u/Otherwise_Nobody_721
10. r/artificial post by u/WAMFT
11. X post by @a16z
12. r/deeplearning post by u/No-Conclusion3720
13. X post by @a16z
14. r/Futurology post by u/Gari_305
15. r/Futurology comment by u/Gari_305
16. X post by @martin_casado
17. X post by @martin_casado
18. r/SaaS post by u/statecs
19. Why companies are becoming a series of loops | Anish Acharya (a16z)
20. The “Owning” Phase of AI Part 1: What Investors Should Look For In Companies Making Their Own Models.