a16z's $1.1B Machine Age Fund Bets the Bottleneck Is Below the Model

VC Tech Radar

2026-08-29

a16z's $1.1B Machine Age Fund Bets the Bottleneck Is Below the Model

By VC Tech Radar • August 29, 2026

The clearest capital-allocation move in this period is a16z's dedicated push into chips, power, and embodied infrastructure. Alongside it, robotics task acquisition, multilingual developer tools, and the fraying boundary between model labs and application harnesses show where the next investable constraints are forming.

1. Funding & Deals

a16z has turned the physical-AI thesis into a dedicated $1.1 billion vehicle. The Machine Age Fund will back founders rebuilding chips, memory, networking, systems software, power, and the machines that bring AI into the physical world. Its partners' stated thesis is that model progress is outrunning the memory, interconnect, power, and cooling beneath it, pushing strong teams from pure software into complex hardware. [1]

The commitment is also organizational: a16z says hardware startups now represent more than 20% of its deal flow, names recent investments including Unconventional AI, Nexthop, Volta, Atoms, and Mind Robotics, and has assembled a team that includes former Intel Data Center Group CTO Guido Appenzeller and data-center veterans Raghu Raghuram and Martin Casado. [2] The managers describe the target founder as a systems builder who can design the chip or system while planning manufacturing, suppliers, and the surrounding ecosystem; they also say first rounds can reach hundreds of millions before a product exists. [3]

2. Emerging Teams

Maiboli is a concrete early signal for cross-language developer tooling. Its builder created a free, MIT-licensed Mac and Windows app that accepts speech in 55-plus languages or mixed-language streams and inserts concise English at the cursor, with an optional rewrite layer for organizing nonlinear speech. The builder reports five weeks of daily use by 20 teammates, more than 3,000 dictations on Gemini 3.5 Flash, and a total bill below 2,500—about one cent per dictation. [4] The founder's background is also notable for the product-led wedge: "an accountant who moved into IT," building around a personal workflow problem rather than a conventional developer-tool pedigree. [4] This is usage evidence, not yet paid-market validation, but it is a useful example of multilingual friction becoming a bottom-up software product.

Sentrint pairs a narrow AI-security workflow with a noisy but promising early funnel. The solo founder reports that three weeks after launch the product had 1,200 visitors, 17 signups, and six paying users; more than 60% of traffic came from Reddit, producing a reported 35% signup-to-paid rate but only a 1.4% visit-to-signup rate. [5] The technical wedge is a three-layer pipeline: multiple security tools produce serialized findings, Claude sees the findings rather than the full codebase to reduce false positives, and a final Claude step creates prompts tailored to the user's chosen LLM platform; scans run in ephemeral instances. [6] The conversion figures should be treated as directional: a commenter notes that six of 17 is too small to establish pricing fit, while Reddit-heavy traffic can inflate paid conversion and crawlers can distort the visitor denominator. [7, 8]

Foundera is a pedigree-led watchlist item rather than a traction case. Founder Cem describes himself as a fifth-time founder who exited a previous company, has built startups since 2011, and has managed accelerator programs including Startupbootcamp. Foundera's agents are intended to validate problems, research markets and competitors, challenge assumptions, define an MVP, connect founders, track milestones, and improve investor visibility. [9] The product is still being built and is seeking early users, so the current signal is founder experience and category ambition—not demonstrated adoption. [9]

3. AI & Tech Breakthroughs

Skild AI's S1 is a potentially important robotics task-acquisition milestone, but the evidence is still partner-reported. A Lightwork segment says the system can receive one video of a person performing a task and execute it without task-specific programming, training, or teleoperation. The claimed change is economic as much as technical: conventional task acquisition required 50–100 hours of puppeteering data and weeks of engineering, while S1 is described as ingesting a demonstration in roughly 11 minutes, adapting to substitute tools, and retrying after failures. [10]

The same segment supplies the right diligence counterweight. Robots that beat

sprinting records still struggle with opening jars, folding socks, stacking cardboard, force modulation, bimanual coordination, peg-in-hole insertion, and long-horizon recovery. For robotics investors, recovery from an unfamiliar physical state is a more informative generalization test than a single athletic benchmark. [10]

LlamaParse is moving enterprise spreadsheet extraction toward structure-aware agents. Its native mode treats spreadsheets as irregular, linked objects with arbitrary rows and columns and cross-sheet dependencies rather than as flat pages; the product uses a tuned model-plus-harness for schema-guided extraction, including dense sheets such as balance sheets. [11] LlamaIndex's accompanying explanation identifies the price/performance problem: the agent must ingest a large, complex interface and emit a large volume of accurate structured output. [12] The investable layer is therefore not simply better OCR, but the harness, schema, and cost controls needed to turn messy enterprise data into reliable machine-readable state.

Evaluation infrastructure is beginning to optimize for statistical stopping. The optstop announcement targets frontier evaluations that can consume hundreds of millions of tokens, stopping trials once estimates are sufficiently precise while continuing runs where uncertainty remains. It is a small but relevant shift from treating every benchmark trial equally toward allocating evaluation budget where it changes the conclusion. [13]

4. Market Signals

AI infrastructure demand is becoming a power, memory, and supply-chain market—not just a model market. An a16z panel cites roughly $700 billion of collective hyperscaler capex this year and says it could reach $1 trillion next year; it describes supply across key components as effectively booked through 2028, some GPUs being resold at four times their purchase price, and simultaneous shortages of power, cooling, memory, and GPUs. [3] The panel's illustrative economics are why specialized infrastructure is attracting capital: if a frontier model costs $3–5 billion to train and must generate about $10 billion in inference revenue, a 20% efficiency improvement could represent $2 billion—enough, in its example, to justify a custom ASIC. [3] Execution remains materially harder than software scaling: permits, grid access, transformers, turbines, and political constraints are already pushing some GPU-seeking companies outside the United States. [3]

Model access is becoming a strategic dependency. OpenAI says it is ending its partnership with Cursor following Cursor's acquisition by SpaceX, with Cursor's direct access to OpenAI models proposed to end on November 12. OpenAI says users can continue using their own API keys and its IDE extensions for Cursor. [14, 15] Harrison Chase's interpretation is that model labs will build strong model-specific harnesses while blocking access from competing labs, leaving cross-model harnesses to independent providers. [16] Jerry Liu makes

the application-layer version of the same argument: companies outside frontier labs will want a mixture of proprietary and open-weight models to optimize performance and margin without being beholden to one provider. [17] LangChain's early support for the new MCP specification is a concrete ecosystem response. [18] The opportunity is a neutral routing and execution layer, though its durability depends on retaining access across increasingly competitive model platforms.

Agent adoption is entering systems of record while stressing their economics and control boundaries. Linear says agents are installed in 95% of paid workspaces and that the share of work they create rose from 3% a year ago to 50%; issues with a pull request attached grew sevenfold since the start of 2026. The same analysis cautions that "share of work" is not "50% of issues," weighting is unspecified, installation is not engagement, and leading AI companies are overrepresented in the cohort. [19]

A separate B2B + AI operator reports agents writing roughly 40GB—about 21 million records—into Salesforce in 30 days, 99% through the API, triggering storage overages after the business had barely used the UI. ServiceTitan then cut off Podium after nine years and roughly 1,000 shared customers when Podium's agent expanded from lead handoff into customer conversations, scheduling, job tracking, and holding the customer record. [20] The pattern is strategic: an agent can make an incumbent system much more useful while also making it easier to replace, and pricing built for human-scale interaction can become punitive when agents write continuously. [20] Reliability is not solved either—the same operator says a renewal agent still occasionally invented numbers after four explicit instructions, requiring human-controlled follow-up. [20]

5. Worth Your Time

- **Watch — Robots Outrun Usain Bolt, Skild's ChatGPT Moment & Google Bids $10M for Spirit Data.** The useful combination is the S1 one-video task-acquisition claim and the sharper benchmark discussion showing why mundane manipulation and recovery matter more than

sprinting. [10]
Robots Outrun Usain Bolt, Skild's ChatGPT Moment & Google Bids $10M for Spirit Data | Lightwork (12:52)

- **Watch — Why Top Founders Are Racing Into AI Infrastructure.** The capex, component-shortage, and custom-ASIC sections provide the clearest current framing of why infrastructure is becoming a first-principles company-building opportunity. [3]
- **Read — The Agents #013.** Concrete cases of agent-written records, platform conflict, workflow speed, and unreliable unsupervised output. [20]
- **Read — LlamaParse's spreadsheet extraction announcement.** A concise product-level example of structure-aware document infrastructure replacing flat extraction. [11, 12]
- **Thread — OpenAI's decision on Cursor.** A timely case study in how model-lab strategy can reprice the risk of building an application or harness on someone else's models. [15, 16]

Sources

1. X post by @a16z
2. X article by @a16z
3. Why Top Founders Are Racing Into AI Infrastructure
4. r/SideProject post by u/Dev14101989

5. r/SaaS post by u/Ok-Okra-3961
6. r/SaaS comment by u/Ok-Okra-3961
7. r/SaaS comment by u/oandresimoes
8. r/SaaS comment by u/oandresimoes
9. r/SaaS post by u/cemnahit
10. Robots Outrun Usain Bolt, Skild's ChatGPT Moment & Google Bids $10M for Spirit Data | Lightwork
11. X post by @jerryjliu0
12. X post by @jerryjliu0
13. X post by @AISecurityInst
14. X post by @OpenAI
15. X post by @thsottiaux
16. X post by @hwchase17
17. X post by @jerryjliu0
18. X post by @sydneyrunkle
19. 50% of the Work Created in Linear Is Now Created by Agents. A Year Ago It Was 3%.
20. When Agents Take Over the System of Record: 40GB Nobody Typed, a Renewal Agent Built in Half a Day, and Why Our Agents Love Clay: The Agents #013