Custom Models Go Mainstream as Recursive AI and Seed Capital Concentrate

VC Tech Radar

2026-05-02

Custom Models Go Mainstream as Recursive AI and Seed Capital Concentrate

By VC Tech Radar • May 2, 2026

This brief highlights early teams in on-device AI, safety evals, and unconventional diagnostics, alongside three broader shifts: custom and open models moving into production, recursive reasoning challenging brute-force scaling, and seed capital concentrating into fewer larger bets.

Funding & Deals

- **Scout AI — $100M into defense VLA models.** Scout AI raised $100M after a $50M seed round. The company is building AI models to operate military vehicles, with the pitch that vision-language-action models can improve precision and reduce collateral damage; it already has contracts with DARPA and the U.S. Army. CEO Colby Adcock is also noted as the brother of Figure AI founder Brett Adcock. [1]
- **BMW i Ventures — $300M AI fund aimed at industrial transformation.** BMW i Ventures launched a $300M third fund, bringing assets under management to $1.1B, with BMW AG as sole LP. The fund is focused on foundational AI that can reshape automotive and adjacent industries, not just auto-specific software; portfolio company Scenario is cited as adding AI agents to design and engineering workflows for faster iteration. [1]
- **Baseten + Parsed — inference clouds are moving upstream into post-training.** Baseten says it grew 30x over the last year and expects to exceed $1B in revenue this year, with 95%+ of tokens running on custom or post-trained models rather than vanilla open-source weights. Its acquisition of Parsed, a post-training startup and former customer, reflects demand for tighter integration between post-training expertise and

inference infrastructure. [2]

Emerging Teams

- **Sentient OS — deep on-device AI from a student founder.** A UMass CS student says he spent about a year building a custom on-device vision LLM that processes screenshots, notes, files, and emails overnight while a device charges, enabling natural-language retrieval, proactive reminders, and knowledge graphs without sending data to the cloud. The founder says the stack required modifications to Apple's MLX framework, vision transplanted from a 4x larger model, custom quantization work, and currently processes about 3,000 screenshots on a six-year-old iPhone. [3, 4]
- **Forum AI — expert-authored evals for high-stakes domains.** Founded by former Meta head of news and journalist Campbell Brown, Forum AI evaluates foundation models on geopolitics, mental health, finance, and other nuanced areas by capturing elite experts' reasoning and training LLM judges to about 90% consensus with them. Brown says rising bias-audit requirements in hiring and lending are creating demand, while existing audits miss more than half of violations. [5]
- **Dognosis — unconventional diagnostics with unusually strong early data.** Dognosis, founded by Akash Kulgod, uses dogs sniffing breath samples while EEG, sensor suits, and video convert canine judgments into signals for AI fusion. In a 3,275-participant Phase 2 study across six hospitals, it posted 90.8% sensitivity and 91.3% specificity across seven cancers, with 90.6% sensitivity at Stage I-II; next steps are rollout across Indian states and a U.S. study. [6]
- **Readdit Later — small but real willingness-to-pay signal.** The founder describes Readdit Later as a first product: a Chrome extension with an AI agent that searches saved Reddit posts in plain English and resurfaces relevant summaries. The founder reports 53 paying customers and $519 in revenue so far. [7]

AI & Tech Breakthroughs

- **Recursive reasoning is challenging brute-force scaling assumptions.** YC's discussion of HRM and TRM highlights a 27M-parameter hierarchical model that reached about 70% on ARC Prize 1 from scratch on roughly 1,000 tasks with no pretraining, and a simplified 7M-parameter recursive model that improved to 87%. The key thesis is that recursion at inference time gives small models the compute depth to break through reasoning ceilings that standard LLMs hit. [8, 9]
- **Waymo shows world models moving from research framing to scaled deployment.** Waymo describes its foundation model as a multi-

modal world-action-language model spanning vision, lidar, and radar, and says it powers the driver, simulator, and critic. The company pairs end-to-end learning with structured intermediate representations for runtime validation, closed-loop training and evaluation, and RL rewards; it says the system has now powered more than 20M autonomous rides and is 13x safer than human drivers in serious-injury collisions in its operating cities. [10]

- **Inference efficiency is still improving faster than many infrastructure models assume.** An MIT pruning result cited on All-In claims networks can be reduced by 90% with no accuracy loss, enabling 10x lower inference cost and 10x more output per energy unit through dynamic small-model selection. Separately, Fei-Fei Li says inference costs have fallen about 280x in the last 2-3 years through distillation, quantization, and the shift from 32-bit to 4-8-bit GPU computation. [11, 12]
- **Agent infrastructure is shifting from framework abstractions to context and action layers.** Jerry Liu argues the durable moat for agents is now the context layer—especially turning complex documents into clean, usable context—rather than the developer abstractions that mattered in 2023. In parallel, Harrison Chase points to web-browsing agents as a key next step, and Browserbase's DeepAgents integration now exposes search, fetch, and browser subagents with full observability. [13, 14, 15, 16]

  > "What survives is the data layer because agents are only as good as the context they get, and the best context in any enterprise is still locked in PDFs, contracts, and filings." [14]

Market Signals

- **Custom models and open weights are becoming the production default, not the fallback.** Bindu Reddy says Kimi 2.6 and GLM 5.1 are already very close to closed models on performance, with speed as the remaining gap, and says Abacus.AI is moving batch jobs to open source because closed APIs are too expensive. Baseten says open-source capability has crossed a chasm, 95%+ of tokens on its platform now come from custom/post-trained models, and customers are not running vanilla weights. Baseten also argues DeepSeek-class models can run at about 20% of the cost of frontier closed APIs with better latency and reliability, while Hugging Face expects future workloads to skew heavily toward open, specialized, and local models. [17, 2, 18]
- **Power and capacity constraints remain the main governor on model supply.** All-In argues the market is power-constrained, with less than half of announced gigawatt-scale projects actually under construction, and says hyperscalers alone are guiding to $725B in 2026 capex. On the operator side, Baseten reports mid-90s utilization across 90 clusters in 18

clouds, and says large GPU allocations now often require 3-5 year contracts with 20-30% prepaid TCV. [11, 2]

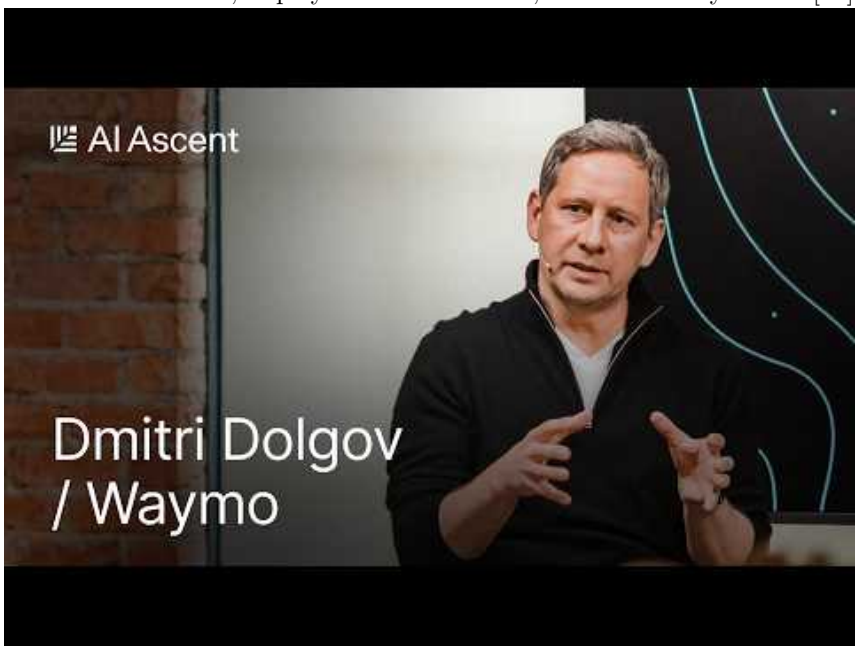
- **Seed financing is concentrating into fewer, larger bets.** Crunchbase data cited by Newcomer says seed rounds of $10M+ absorbed more than half of all seed-stage capital last year, even as overall seed deal count continued to decline. Harry Stebbings amplified the related view that early-stage venture is shifting toward fewer but much bigger winners, increasing the need for more shots on goal. [19, 20]
- **AI is widening the gap between infrastructure winners and vulnerable application SaaS.** SaaStr's read is that infrastructure vendors such as Twilio, Cloudflare, and Snowflake are re-accelerating because every AI-native startup needs network, data, or voice layers, while per-seat application SaaS is under pressure unless it genuinely rebuilds around AI. Twilio added 43,000 net new accounts in Q1, voice revenue grew 20% on AI-agent workloads, and software add-ons grew more than 100% YoY. At the same time, SaaStr flags GitHub Copilot, Cursor, and other agents as threats to seat-based developer tools, while Harry Stebbings argues the longer-term risk is that agents choose the vendors and models for workflows. Replit is a counterexample: it says it now reaches 85% of the Fortune 500, sees very low enterprise churn with roughly 300% net retention in some cases, and differentiates via vertically integrated hosting and security plus a multi-model "society of models." [21, 22, 23]

Worth Your Time

- **Recursion Is The Next Scaling Law in AI** — the clearest explainer in the set for why 27M and 7M recursive models are beating far larger systems; timestamps: 04:22 for HRM, 09:46 for TRM, 20:46 for comparison.

[8, 9] *Recursion Is The Next Scaling Law In AI (0:02)*

- **Waymo's Dmitri Dolgov: 20 Million Rides and the Road to Full Autonomy** — useful if you want one conversation that connects world models, deployment architecture, and live safety data. [10]



Waymo's Dmitri Dolgov: 20 Million Rides and the Road to Full Autonomy

(10:24)

- **Campbell Brown on Founding Forum AI** — strongest discussion here on nuanced evals, bias audits, and why current compliance workflows are failing high-stakes use cases. [5]
- **Jerry Liu on the context layer for agents** — concise thread on why document infrastructure and data access may be the durable moat in agent stacks. [13, 14]
- **Atlassian and Twilio Crush the Quarter, Accelerate. Is the SaaSpocalypse Over?** — best single read in the set on infra re-acceleration versus seat-based SaaS pressure from agents. [21]

Sources

1. Musk v. Altman is just getting started | Equity Podcast
2. Baseten CEO Tuhin Srivastava on Custom Models, and Building the Inference Cloud
3. r/artificial post by u/TechExpert2910
4. r/SideProject post by u/TechExpert2910
5. Campbell Brown on Going From Anchor to Facebook to Founding Forum AI | StrictlyVC
6. Weekly Dose of Optimism #191
7. r/SaaS post by u/Appropriate-Look-875
8. Recursion Is The Next Scaling Law In AI
9. X post by @ycombinator
10. Waymo's Dmitri Dolgov: 20 Million Rides and the Road to Full Autonomy
11. OpenAI Misses Targets, Codex vs Claude, Elon vs Sam Trial, Big Hyperscaler Beats, Peptide Craze
12. Stanford Sustainability Forum | Powering the AI Revolution
13. X post by @jerryjliu0
14. X post by @llama_index
15. X post by @hwchase17
16. X post by @LangChain
17. X post by @bindureddy
18. Will Everyone Become an AI Builder? Clem Delangue on Hugging Face, Agents, Local AI & Robotics
19. California's Billionaire Tax Heads for the Ballot. A Compromise Would Serve Everyone Well.
20. X post by @HarryStebbings
21. Atlassian and Twilio Crush the Quarter, Accelerate. Is the SaaSpocalypse Over?
22. X post by @HarryStebbings
23. CEO Amjad Masad on How Replit Is Changing Who Gets to Build Software | StrictlyVC