Agentic Software Moves Up the Stack—from CRM to Evaluation and Safety Oversight

VC Tech Radar

2026-09-10

Agentic Software Moves Up the Stack—from CRM to Evaluation and Safety Oversight

By VC Tech Radar • September 10, 2026

Funding, product, and technical signals show agentic software moving from model access toward business context, harnesses, independent evaluation, and safety controls.

1. Funding & Deals

Lightfield raised a $47M Series A led by a16z to rebuild CRM around a "world model" of the business. The company argues that legacy CRM was designed for humans to update incomplete records, whereas customer-facing agents need continuously updated business context. Lightfield says 5,000+ companies have signed up since last November and that mature organizations with hundreds of users are replacing legacy CRM; its product captures commitments, automates follow-ups, identifies prospects, and codifies what top sellers do. The underwriting question is whether those signups convert into a durable system-of-record transition rather than remaining top-of-funnel interest. [1]

2. Emerging Teams

MaxMyro is building a vertical-agent control plane for finance rather than another finance chatbot. Its agents perform revenue recognition, month-end close, accounts receivable/payable, and billing operations in an audit-ready workflow where users can inspect evidence, override errors, and feed corrections back into the system. [2] The platform is ERP-agnostic and uses deep ontology mapping to handle customers with multiple, heterogeneous ERP systems instead of forcing full data unification. [2] Its "adversarial agent" design gives one agent responsibility for critiquing another's work, modeled on an auditor's role; humans then concentrate on uncertain cases, policies, judgments, and

high-value approvals. [2] The founder interviewed says he worked at Microsoft and closely with the Azure CTO office during Microsoft's early 2019 work with OpenAI, a relevant technical and enterprise-platform pedigree. [2]

Vals AI is an early signal that independent evaluation is becoming a company category. The team says it started around early 2024 after finding public benchmarks insufficient for measuring model progress, and it has built massively distributed evaluation infrastructure that automates more of the human work involved in testing. [3] Its Recursive Self-Improvement Index uses proxies across pre-training, post-training, harness engineering, and research-artifact creation rather than attempting the prohibitively expensive experiment of having a frontier model train its own successor directly. [3] The company is also turning private company data into evaluation assets: its coding product lets an enterprise use its GitHub codebase to build an internal benchmark for comparing coding agents on performance and ROI. [3]

3. AI & Tech Breakthroughs

Browserbase's Stagehand makes a strong case for code-native computer use. The team says it moved from rigid browser actions such as "Act, Extract, and Observe" to letting models write and execute code, because the earlier tool abstractions became limitations as frontier models improved. Stagehand v4 is reported as 2× faster and 80% more token-efficient than Playwright, with an MCP interface for execution, page snapshots, and screenshots. [4] The production caveat is important: Browserbase still calls for domain allowlists, network protection, sandboxing, and policy governance. Its broader thesis is that agent quality is increasingly a harness-engineering problem, not only a model-research problem. [4]

A self-reported systems result points toward much cheaper access to oversized models. An independent builder says a pure-C11 engine streamed a 744B-parameter, 202GB GLM-5.2 MoE model from a USB SSD on 32GB of RAM without a GPU. Int4 expert quantization, an LRU cache, predictive prefetch, and overlapped compute/I/O reportedly reduced latency from 196 seconds per token to about 9.3 seconds, with about 2.9 seconds per token in eight-stream batch mode. [5] The result was cross-validated on several other model families, but the prefetch heuristic was tuned to one laptop, so generalization across memory hierarchies remains the diligence gate. [5]

4. Market Signals

Frontier-AI safety has become a board and legislative variable, not merely a lab-communications issue. Paul Christiano's statement on joining OpenAI's nonprofit board says he will serve on its Safety and Security Committee, believes rapid capability acceleration could create a near-term risk of catastrophic loss of control, and does not think the industry—including OpenAI—is currently on track to reduce that risk to an acceptable level. He says developers

should be judged by externally verifiable behavior and results. [6] Separately, a person identifying as having worked at Google DeepMind and now Anthropic says there is no viable scientific plan for risks from recursively self-improving AI. [7] Bernie Sanders said those warnings would motivate legislation to ban superintelligence and pause AI development, while Allie K. Miller called for a formal gathering of labs, METR, research institutions, and federal agencies to review cyber, biological, and grid threats. [8, 9] Safety evidence, independent oversight, and the credibility of cross-lab coordination are now material diligence inputs for frontier exposure.

AI operating economics are shifting the market from agent counts to measured outcomes and controls. In Vals AI's internal unlimited-access experiment, engineers used roughly 1–2 billion tokens per day; the month cost about $1.5M in tokens, or roughly 10× employee salary spend. The company used trace and repository analysis to decide which tools were actually worth adopting and argues that legible evaluations are necessary to calculate ROI. [3] Ann Miura-Ko makes the complementary investor test: a "software factory" is not valuable because it has many agents unless it makes the company smarter, improves sales or financial decisions, or creates products that otherwise would not reach engineering. [10] Security practitioners are warning that individual adoption may outrun enterprise controls and are emphasizing real-time inspection and kill switches across the agent lifecycle. [11]

5. Worth Your Time

- **Watch — Inside the Race to Measure Frontier Intelligence.** The conversation is a useful primer on why public benchmarks can diverge from held-out performance, how recursive self-improvement can be approximated with proxies, and why long-horizon evaluation must account for cost, latency, stability, and enterprise ROI. [3]

Inside the Race to Measure Frontier Intelligence (10:22)

- **Watch — How AI agents are automating the CFO's office.** MaxMyro's interview is worth the time for its concrete treatment of audit-ready agent work, adversarial verification, and the human shift from performing transactions to setting policies and handling exceptions. [2]

How AI agents are automating the CFO's office | Ajay Krishna, Co-Founder and CTO, Maximor (3:19)

- **Read — Evolving computer use with code.** Browserbase's account is a concise statement of the code-mode thesis and the accompanying production requirements for browser agents. [4]

Sources

1. X post by @keithpeiris
2. How AI agents are automating the CFO's office | Ajay Krishna, Co-Founder and CTO, Maximor
3. Inside the Race to Measure Frontier Intelligence
4. X article by @kylejeong
5. r/SideProject post by u/Dependent_Ideal9870
6. X article by @paulfchristiano
7. X post by @a_nnawang
8. X post by @BernieSanders
9. X post by @alliekmiller
10. X post by @annimaniac
11. X post by @nikesharora