Agent Benchmarks Rise as Context Infrastructure and Physical AI Take Shape

VC Tech Radar

2026-05-12

Agent Benchmarks Rise as Context Infrastructure and Physical AI Take Shape

By VC Tech Radar • May 12, 2026

Cognition and Legora set the strongest traction markers in this cycle, while new startups target agent analytics, context delivery, and secure local workflows. The broader investor backdrop is widening toward physical AI, voice, and defense autonomy as talent and chip economics become less forgiving.

1) Funding & Deals

- **Cognition AI:** Cognition is reportedly raising at around a $25B valuation after Devin reached a $445M revenue run rate in its first 18 months, with usage doubling every eight weeks and customers including the US Army, Goldman Sachs, and Mercedes-Benz [1]. The company was founded in November 2023 around the view that AI agents would work in the background like 24/7 coworkers, and it shipped Devin in March 2024 despite early public criticism [1].
- **VoriHQ:** Vori raised a $22M Series B around a clear retail thesis: grocery stores do more volume than restaurants and hotels, yet most still run on clipboards. Its product is a modern operating system for grocers, including AI agents that automatically update shelf items when costs change [2].
- **Monaco:** Sam Blond's Monaco reportedly closed millions in revenue days after launch by using forward-deployed account executives who configure the AI SDR live for customers. In this model, the sale and deployment happen at the same time, reducing reliance on long evaluations and discounting [3].

2) Emerging Teams

- **Legora:** Legora has become one of the clearest legal AI traction stories: $100M ARR in 18 months, onboarding 50 customers every 14 days, and a reported $50M qualified pipeline from a Jude Law campaign [4]. Its operating model is notable: former attorneys serve as Legal Engineers alongside forward-deployed engineers, and the company treats change management as part of the product [4]. Paul Graham called it the most impressive startup he has visited in years [4].
- **Scope:** Scope is building agent analytics for the agent era. It runs real workflows across Claude Code, Codex, Cursor, and similar tools so companies can see when agents choose them, get stuck, or pick competitors, and what to change [5]. YC's launch post identified founder @anandPa94 [5].
- **Weavable:** Weavable is an MCP-native context layer that preprocesses data from HubSpot, Jira, Slack, Zendesk, Notion, and more before it reaches an agent [6]. The team reports 85% favorable results in LLM-as-a-judge evaluations against baseline retrieval and up to 90% token savings, using an evolving changelog across systems for entity resolution, freshness checks, and ranking [6, 7].
- **Framewise Health:** Framewise Health turns medical records, institutional protocols, and drug data into personalized videos for patient onboarding, adherence, and recovery [8]. YC's launch post named founders @tanekimm and @sourdoggy8 [8].

3) AI & Tech Breakthroughs

- **Laptop-scale open weights are improving faster than laptop hardware.** On unchanged 128 GB MacBook Pro hardware, the best open-weight model runnable locally improved from score 10 to 47 on the Artificial Analysis Intelligence Index between May 2024 and May 2026, a 4.7x gain and a doubling every 10.7 months [9].
- **Parallel agent orchestration is becoming a product category.** Replit launched Parallel Agents, which lets users run up to 10 agents in parallel, each with its own copy of the app and its own computer, then merge results agentically. Amjad Masad described the breakthrough not as multiple agents by itself, but correct orchestration and seamless merge-back, with projects moving 10x faster [10, 11].
- **Secure local-first file handling is turning into core agent infrastructure.** LlamaIndex released sandboxed-lit, a Rust CLI agent for parsing PDFs, images, and Office files with LiteParse, a secure sandbox powered by microsandbox, full filesystem mounting, and Bash access [12]. Jerry Liu framed agents plus file sandboxes as a 2026 trend [13].

- **RL fine-tuning has a new efficiency lever for long prompts.** A prompt-caching approach for RL fine-tuning computes the prompt once across grouped responses while preserving gradient flow, producing 5x to 7.5x speedups on long-prompt, short-response workloads in reported Qwen3.5-4B benchmarks [14].
- **Compiler experimentation is getting much more accessible.** A hackable ML compiler built in 5,000 lines of Python lowers small models through six IRs to CUDA and, on RTX 5090 FP32, reports geomean performance of 1.11x versus PyTorch eager and 1.20x versus torch.compile, with wins up to 4.7x on some operations [15].

4) Market Signals

- **OpenAI tender liquidity is becoming a funding source for the next wave.** SaaStr estimates cumulative OpenAI tenders since 2021 have likely created 300 to 500+ employees with more than $10M in realized secondary cash, and argues hundreds are already angel-investing into new B2B + AI startups in San Francisco [16]. The same piece says top AI/ML talent now expects $500K+ base plus liquid equity, with secondary tenders every 12 to 18 months increasingly treated as table stakes [16]. It also argues that non-founder operators at the right labs can now outperform typical founder economics [16].
- **Physical AI is moving from side thesis to investable category.** Fei-Fei Li says the market is saturated with language AI use cases while underappreciating how perceptual and physical work really is [17]. World Labs is building large world models aimed at understanding and navigating physical space, while Chelsea Finn's Physical Intelligence is building a foundation model for robotics [17, 18].
- **Voice AI still looks early, but enterprise traction is real.** Investors at the Cerebral Valley Voice Summit described the category as still in a Copilot era, with enterprise ahead of consumer [19]. At the same time, Assort Health says its voice agents have already handled 150 million patient interactions across 5,000 providers, Abridge sees healthcare regulation and privacy as a moat, and OpenAI's newer realtime voice models can reason mid-conversation [19]. Deepgram predicted a five-minute voice Turing Test could be passed by year-end through better context memory [19].
- **Semiconductor economics are no longer a free tailwind.** Exponential View flagged Bloomberg reporting that TSMC does not plan to use ASML's High-NA EUV tool through 2029 because of cost [20]. The same essay notes that cost per transistor stopped falling in 2011, and that even lithography-driven cost improvements have now reversed, which matters because modern AI infrastructure was built on generations of cheaper chips every 18 to 24 months [20].

- **Defense autonomy has shifted from taboo to urgent.** a16z argues the US has the talent advantage but is losing the production race in autonomous systems, and frames the stakes as whether the country reaches the next conflict with overwhelming autonomy superiority or cedes that advantage to adversaries [21]. Separately, My First Million described a cheap-drone problem where $2M missiles are used against $200 drones and argued that the next startup waves are increasingly clustering around AI labs, defense tech, and hard tech or robotics [22].

5) Worth Your Time

Fei-Fei Li on large world models

Watch for a concise articulation of why physical and perceptual intelligence may be the next major AI platform shift, and how world models could reconstruct, predict, and simulate physical space [17].

AI Is Moving Beyond the Screen. Are CEOs Ready? (0:51)

Marc Andreessen on the builder era

Watch for the strongest current argument that AI is creating a new builder role, with leading-edge programmers reportedly becoming 20x more productive and firms actively seeking AI-native talent [23].

The Golden Age Thesis | Marc Andreessen on MTS (26:47)

The broken bargain of Moore's Law

Read for a useful framing of why TSMC's hesitation on High-NA matters for AI investors who have assumed another decade of automatic compute cost declines [20].

Newcomer's voice summit roundup

Read for operator-level traction and infrastructure details across Wispr Flow, Assort Health, Abridge, Cartesia, LiveKit, and Deepgram [19].

Harry Stebbings on Legora's enterprise AI playbook

Thread for a practical breakdown of why brand can create pipeline, but expert services and change management still determine whether enterprise AI workflows actually stick [4].

Sources

1. X post by @colossusmag
2. X post by @ycombinator

3. Stop Discounting. Start Deploying. The New Way to Close Deals in B2B + AI.
4. X post by @HarryStebbings
5. X post by @ycombinator
6. r/SideProject post by u/abeshius
7. r/SideProject comment by u/abeshius
8. X post by @ycombinator
9. X post by @ClementDelangue
10. X post by @Replit
11. X post by @amasad
12. X post by @llama_index
13. X post by @jerryjliu0
14. r/deeplearning post by u/girishkumama
15. r/MachineLearning post by u/NoVibeCoding
16. OpenAI Has Already Created 300+ Decamillionaires. More Than a Decade of B2B IPOs Combined. Before Going Public.
17. AI Is Moving Beyond the Screen. Are CEOs Ready?
18. X post by @ycombinator
19. 13 Videos From the Cerebral Valley Voice Summit: Sierra's Bret Taylor, Wispr Flow's Tanay Kothari, MiniMax's Linda Sheng & More
20. The broken bargain of Moore's Law
21. X post by @a16z
22. I put 80% of my money in the S&P after a billionaire investor told me not to
23. The Golden Age Thesis | Marc Andreessen on MTS