Compute Crunch, Agent Infrastructure, and the Case for Automated AI R&D

VC Tech Radar

2026-05-05

Compute Crunch, Agent Infrastructure, and the Case for Automated AI R&D

By VC Tech Radar • May 5, 2026

The clearest signals this cycle sit at the intersection of compute scarcity, agent infrastructure, and early evidence that AI is taking over more of the software and research stack. This brief highlights the limited direct deal activity, the emerging teams with real traction or technical edge, and the market shifts most relevant to AI investors.

1) Funding & Deals

- **Jason Calacanis is making a direct bet on alternative AI compute markets.** He said he has invested $750k * *in * *$**TAO** and Bittensor subnets because he expects a few home runs from the entrepreneurial energy in those networks. He highlighted **Chutes** (Subnet 64) as the largest subnet by market cap, with **1,049 GPUs** and pricing designed to intensify competition [1].
- **Sequoia also surfaced a seed-stage signal.** Sequoia partner **Andrew Reed** posted *neocloud seeding*, a sparse but direct indication of a new seed investment [2].

2) Emerging Teams

- **A former enterprise CPO is turning MCP pain into a company.** While leading product at a supply-chain carbon SaaS used by **Adidas, Swarovski, and Puma**, the founder said customers were asking for Claude-like agents to use the product directly, not just for new AI features. He left after concluding that discoverable schemas, agent-scoped auth, per-customer scoping, rate limits, observability, and tool descriptions amounted to a second stack; his implementation pattern

auto-generated **173 tools** from tagged routes while preserving OAuth and RBAC [3, 4, 5].

- **Expanse is a timely YC compute-infrastructure launch.** YC says the product unlocks wasted GPU capacity by matching jobs to the right resources, speeding runs, and debugging failures across **cloud and on-prem HPC** [6].
- **Kuli already has real enterprise deployment at launch.** YC describes it as an AI coworker for marketers that watches social video to find trends and plans and runs creator campaigns; it says the product is already live at **Fortune 100 brands** and makes teams **10x** more efficient [7].
- **RepoInspect is building security specifically for agentic software.** The solo founder, an AI engineer with **9 years** in the space, built a two-pass system: deterministic AST taint tracking to find hotspots, then an AI agent to verify exploitability. He says it found multiple bugs in popular AI frameworks and now supports local LLMs for private audits [8].
- **ClankerRank has early traction in a broken hiring market.** The assessment platform says **550** users across **15 countries** arrived organically to solve real RAG, agent-design, and prompt-chain debugging tasks; its core claim is that the top **10%** stand out on **judgment**, not on resume keywords or standard coding screens [9].

3) AI & Tech Breakthroughs

- **Claude Code is showing a frontier-lab version of agentic software development.** Boris Cherny said the model wrote **100%** of the TypeScript/React codebase early on and still writes **100%** of his code today. He described working across **5-10 sessions**, **hundreds** of active agents, and sometimes **thousands** overnight, with new primitives such as recurring **/loop** jobs, server-side **routines**, and sub-agent parallelism [10].
- **Synthetic pretraining is improving reasoning in sub-1B models.** Tufa Labs says synthetic pretraining let **0.8B** models beat baselines on **GSM8K** and **MATH500**, deliver **2-3x** larger few-shot gains, and match baseline performance with **3-6x** fewer training tokens; a same-size generator was enough [11].
- **KV-cache compression is posting strong numbers on modest hardware.** OmniStack-RS reports **3.37x** compression, **0.69 ms** P99 kernel latency, and **1,633.93 queries/sec** on an **NVIDIA A10**, using INT4 quantization, a 1-bit residual, and a fused Triton attention kernel [12].
- **Local inference on Apple silicon keeps moving up-market.** A solo founder said a Mac-native runtime built in **32 days** now runs **Qwen3.5-397B-A17B** at about **1.6 tok/s** on a **64GB Mac Studio** through a paged MoE engine with **14GB** peak RAM; the stack is **Tauri + Rust + MLX**, with **568** tests and no outside funding [13].

4) Market Signals

- **The compute crunch is intensifying.** Exponential View cites **B200** rental prices up **114%** in six weeks, a **6x+** widening premium versus H200, demand from **40** Lightning AI customers for **400,000 GPUs** against a **40,000** GPU fleet, and Microsoft's requirement that Blackwell customers commit to at least **1,000 chips** for a year [14].
- **Automating AI R&D is hardening into an investable thesis.** Jack Clark argues there is a **60%+** chance of no-human-involved AI R&D by end-**2028**, with proof-of-concept possible in **1-2 years**, citing coding benchmark saturation, stronger reproducibility results, longer agentic work horizons, faster training optimization, and early automated alignment-research wins. He also notes that **hundreds of billions** of dollars are being aimed at the category; Jason Calacanis called the forecast *material and realistic* [15, 16].
- **LLM citation optimization is emerging as a distribution wedge.** The founder of **learnwithpath** said ChatGPT sent **782** visits in 30 days versus **308** from Google after he reworked content for LLM citation using quick-answer boxes, JSON-LD schema, tables, and subreddit-derived FAQs [17]. Another founder said **Docsio**, just one month old, became the **#1** recommendation in an incognito ChatGPT query despite no funding, weak domain authority, and no G2 or listicle presence, helped by daily researched long-form content, schema, and aggressive internal linking [18].
- **In both autonomous systems and AI hiring, the bottleneck now looks like judgment.** After **8 months** in production, LocusFounder says capability is no longer the main constraint for autonomous storefronts, copy, and ad management; the dangerous failures are confidently wrong decisions outside expected conditions [19]. ClankerRank reports that the top **10%** of AI engineering candidates are differentiated less by model knowledge than by judgment on whether a problem needs **RAG**, an **agent**, or neither [9].
- **The valuation market is rewarding growth and genuine AI nativeness, not surface-level AI features.** Based on more than **1,002,048** startup valuations run through SaaStr.ai tools, a **$5M ARR** company growing **200% YoY** is worth more than a **$20M ARR** company growing **30% YoY**. SaaStr says true AI-native companies with real revenue get a premium, while *we added an AI feature* gets almost none; it also launched an API Report Card because many B2B APIs are still not usable enough for agent builders [20].

5) Worth Your Time

Performative AI Usage and Pragmatic AI Usage exist in separate, parallel economies. [21]

- **Import AI: AI systems are about to start building themselves** — the best single essay in the set on why automated AI R&D is becoming a

real strategic question, not just a thought experiment [15].

- **Boris Cherny on Claude Code** — useful if you want to see how frontier-lab teams are actually operationalizing agentic coding, especially around loops, routines, and parallel agents [10].

Anthropic's Boris Cherny: Why Coding Is Solved, and What Comes Next (5:41)

- **Dalton Caldwell on performative vs. pragmatic AI usage** — a compact framework for separating hype-driven token burn from durable, ROI-positive adoption [21, 22].
- **Martin Casado on minimal agent harnesses** — a useful thread on how little infrastructure he says agents may need, plus examples of an agent building Discord, browser, WhatsApp, and X support from a minimal setup [23, 24].
- **learnwithpath on optimizing for LLM citations** — one of the clearest tactical writeups in the set on AI-search distribution, with concrete formatting and schema choices that preceded ChatGPT overtaking Google as a referrer [17].

Sources

1. X post by @Jason
2. X post by @andrew___reed
3. r/SaaS post by u/CrewPale9061

4. r/SaaS comment by u/CrewPale9061
5. r/SaaS comment by u/nsjames1
6. X post by @ycombinator
7. X post by @ycombinator
8. r/SaaS post by u/WinterSpecial7970
9. r/SaaS post by u/Equivalent-Device769
10. Anthropic's Boris Cherny: Why Coding Is Solved, and What Comes Next
11. r/deeplearning post by u/m_sap
12. r/deeplearning post by u/Superb_Housing9628
13. r/SideProject post by u/ur_dad_matt
14. Data to start your week: AI boom, nowhere near the ceiling
15. Import AI 455: AI systems are about to start building themselves.
16. X post by @Jason
17. r/SaaS post by u/PlusGap1537
18. r/SaaS post by u/sinatrastan
19. r/artificial post by u/IAmDreTheKid
20. We Just Crossed 1,000,000 Startup Valuations on Our Free SaaStr.ai VC Tools
21. X post by @daltonc
22. X post by @daltonc
23. X post by @martin_casado
24. X post by @martin_casado