Cheaper Inference, Agent Memory, and the New Compute Moat

VC Tech Radar

2026-04-18

Cheaper Inference, Agent Memory, and the New Compute Moat

By VC Tech Radar • April 18, 2026

Capital this cycle centered on cheaper inference and strategic autonomy bets, while early teams pushed forward on agent memory, semantic retrieval, and non-invasive interfaces. The broader read-through is that AI competition is shifting from raw model novelty toward context infrastructure, compute access, and differentiated workflows.

1) Funding & Deals

Disclosed financing in this set clustered around cheaper inference and hardware-linked autonomy rather than pure application SaaS [1].

- **Parasail — $32M for cheaper inference infrastructure.** TechCrunch said Parasail raised about $32M to scale AI inference infrastructure in a cheaper way, landing into a market where firms are increasingly focused on token use and budget efficiency [1].
- **Wave — $60M strategic extension from AMD, Arm, and Qualcomm.** The UK autonomous-driving company added a $60M extension on top of its more than $1B Series D, and TechCrunch said the semiconductor investors are taking real equity, not in-kind credits, to help scale Wave's hardware-agnostic, end-to-end neural-network stack; Uber plans another $300M based on milestones [1].

2) Emerging Teams

- **Sabi — stealth BCI company with top-tier backing.** Not Boring said Sabi emerged from stealth backed by Khosla Ventures, Accel, Initialized, and OpenAI VP Kevin Weil, with a non-invasive cap/beanie that

uses 70,000-100,000 EEG sensors and a Brain Foundation Model aimed at 30 words per minute; shipping is expected at year-end [2].

- **Gaia — college-student founder tackling tool-calling scale.** The solo builder said Gaia hit a wall at 200 tools, then fixed hallucinations and context bloat by embedding tools in ChromaDB and retrieving them semantically at runtime; the system now routes across a three-layer comms/executor/subagent architecture and claims to scale to thousands of tools without degradation [3].
- **AgentID — shared memory for multi-tool workflows.** AgentID is pitched as a shared memory, context, and identity layer so multiple AI tools stop redoing setup and burning tokens; the founder says its Caveman compression layer cuts token usage by up to 65% in some workflows, and early commenters validated the pain around repeated context loss while pushing for harder proof on completion rates and scoped resets [4, 5, 6].
- **AgentMailr — agent-native email infrastructure, already in production.** The founder built it around persistent mailboxes, thread-level routing, sender filtering, and reliable inbound webhooks; shipped features include mailbox provisioning per agent, routing rules, allowlists/blocklists, and BYOS, with follow-up discussion focused on protections against agent loops using sender rules and thread-ID tracking [7, 8].

3) AI & Tech Breakthroughs

- **Structured context is starting to beat brute-force RAG in code and tool use.** One builder reported 80% hit@5 retrieval across 18 repos and 90 tasks using only regex + TF-IDF over function signatures and class shapes, versus a 13.6% random baseline, with a 98.1% token reduction and no embeddings or ML [9]. A related code-memory project, Ix, maps repos into graphs of files, functions, relationships, and dependencies so models query structure instead of chunks [10]. Gaia makes the same bet on tools: semantic retrieval replaced prompt-listed tool search and was said to take the system from dozens of tools to thousands without degradation [3].
- **Persistent runtimes are getting more autonomous.** Springdrift injects a structured self-state block called a sensorium into each cycle, and its author described an episode where the agent noticed a missing writer agent from passive context and rerouted work without a diagnostic tool call [11]. Agent Relay is making a similar infrastructure bet from the other direction: synced files across multi-agent sandboxes and virtual file mounting from systems like Notion, pitched as faster and lower-token than API-heavy access [12, 13].
- **OpenClaw's core insight is UX, not a new base model.** The product is framed as winning on ergonomics because messaging channels like iMessage, WhatsApp, and Telegram make delayed replies feel normal, reducing the pressure for instant-but-shallow responses; Garry Tan separately called OpenClaw "straight magic" and Peter Steinberger's TED talk "a revelation," while All-In said OpenAI has hired Steinberger as it

pushes for the agent platform layer [14, 15, 16, 17].

"that's the magic of openclaw - same underlying tech, different consumer mental model" [14]

- **Coding agents are starting to show real utility in personalized medicine.** Patrick Collison said agents working over his genome surfaced a roughly 30x higher melanoma predisposition and recommended follow-on tests, supplements, and more frequent screening; he estimated the analysis at under $100 on top of a few hundred dollars for sequencing, while noting the agents still need monitoring and re-steering. Marc Andreessen publicly co-signed the use case [18, 19].

4) Market Signals

- **Anthropic's enterprise-coding focus is being cited as a major growth driver.** On All-In, speakers said Anthropic and OpenAI were both around a $30B run rate at the start of Q2, while also noting Anthropic's figure may be lower on an apples-to-apples basis because of channel-partner revenue; the same discussion said Anthropic has been growing roughly 10x/year versus OpenAI's 3-4x/year, with enterprise coding and metered usage explaining the gap. The speakers also said secondary markets now value Anthropic above OpenAI, and that OpenAI is pivoting harder toward business customers and the agent platform layer [17].
- **Compute supply and hardware fit are becoming larger competitive variables.** All-In argued frontier labs have grown to the point where depending on hyperscalers is a strategic mistake, and cited increasing siting resistance, including a Maine ban and claims that roughly 40% of contested data-center projects get canceled [17]. In parallel, Gavin Baker argued model portability is eroding as hardware topologies diverge and labs optimize for inference economics, not just training, which raises switching costs and rewards tighter co-design between models and systems [20].
- **The application moat is moving beyond the wrapper.** Clouded Judgement lays out a progression from thin wrapper to harness to post-training and eventually pre-training proprietary models, arguing that early winners like Cursor are already moving into phases 3-4 [21]. Garry Tan's operating version is "fat skills, fat code, thin harness," and he argues many critiques of agents are really critiques of naked LLM use without tools, deterministic code, or context management [22, 23, 24, 25].
- **This still looks like an expansionary spend cycle, not a mature ROI market.** Clouded Judgement says many companies are currently "over earning" on rapid AI-spend growth and token-maxing behavior, but expects an optimization phase once budgets balloon, which should separate differentiated vendors from rising-tide beneficiaries [21]. Parasail's financing around cheaper inference infrastructure sits inside that theme

[1].

- **Founder supply remains broad, but the skill gap may widen.** Garry Tan pointed to YC funding 800+ mostly first-time founders as evidence that deciding what to build still matters even as tools improve [26]. In a separate post he highlighted, heavy users were described as encoding full workflows in plain-English markdown, with the claim that "engineering context = engineering code" [27, 28].
- **Investors are re-litigating what counts as ARR.** One critic warned that reporting stepped multi-year contracts as current ARR can inflate figures by roughly 3x and mask negative first-year margins from bundled forward-deployed engineers, while Martin Casado pushed back that using exit ARR as current is not that common and is less problematic than other reporting games such as treating GMV as ARR [29, 30].

5) Worth Your Time

- **All-In on Anthropic, OpenAI, and the datacenter constraint** — covers the claims that Anthropic is compounding faster than OpenAI on enterprise coding economics, that OpenAI is pivoting toward business customers and agents, and that frontier labs now need their own infrastructure [17].

OpenAI's Identity Crisis, Datacenter Wars, Market Up on Iran News, Mamdani's First Tax, Swalwell Out (22:29)

- **Gavin Baker's portability thread** — useful on why tokens per watt per dollar now dominate, why co-designed models run worse on the "wrong" hardware, and why U.S./China AI stacks may diverge [20].
- **Clouded Judgement: "Rising Tide, Hidden Risk"** — lays out the case that today's AI spend boom is masking over-earning, and that the next moat may shift from harnesses to proprietary models [21].
- **Gaia's tool-calling writeup** — details the failure mode at 200 tools, the move to semantic retrieval, and the comms/executor/subagent architecture that followed [3].
- **Peter Steinberger's TED talk on OpenClaw** — Garry Tan called it "a revelation," and All-In later noted that OpenAI hired Steinberger as it pushes deeper into the agent platform layer [16, 17].

Sources

1. Are we tokenmaxxing our way to nowhere? | Equity Podcast
2. Weekly Dose of Optimism #189
3. r/SideProject post by u/Ok-Programmer6763
4. r/SideProject post by u/Single-Possession-54
5. r/SideProject comment by u/Glum_Foundation5476
6. r/SideProject comment by u/Kevin_Xiang
7. r/SideProject post by u/kumard3
8. r/SideProject comment by u/kumard3
9. r/SideProject post by u/Independent-Flow3408
10. r/SideProject post by u/PlusLoquat1482
11. r/MachineLearning post by u/s_brady
12. X post by @Khaliqgant
13. X post by @dunkhippo33
14. X post by @illscience
15. X post by @garrytan
16. X post by @garrytan
17. OpenAI's Identity Crisis, Datacenter Wars, Market Up on Iran News, Mamdani's First Tax, Swalwell Out
18. X post by @patrickc
19. X post by @pmarca
20. X post by @GavinSBaker
21. Clouded Judgement 4.17.26 - Rising Tide, Hidden Risk
22. X post by @garrytan
23. X post by @garrytan
24. X post by @garrytan
25. X post by @garrytan
26. X post by @garrytan
27. X post by @garrytan
28. X post by @DoDataThings

29. X post by @scottastevenson
30. X post by @martin_casado