Blank Bio's Seed, Exa's Search Bet, and the Agent-Native Infrastructure Shift

VC Tech Radar

2026-05-21

Blank Bio's Seed, Exa's Search Bet, and the Agent-Native Infrastructure Shift

By VC Tech Radar • May 21, 2026

Blank Bio's seed and Exa's search financing framed the capital signals, while YC launches and new commentary from Baseten, Railway, and Cohere sharpened the investment case around agent-native software, post-training, and compute economics.

Funding & Deals

- **Blank Bio:** Blank Bio raised a $7.2M seed with a strategic collaboration from PacBio. The company is training foundation models on bulk RNA-seq to help pharma design better clinical trials by learning patient heterogeneity and building prognostic and predictive biomarkers from tumor transcriptomes. Announcement [1]
- **Exa:** Exa raised $250M at a $2.2B valuation in a Series C led by a16z. Not seed-stage, but still a clear thesis-confirming financing: Exa is positioning as search infrastructure for AI agents, especially on long-tail, high-alpha queries where traditional engines fail, and a16z says developers and agents are already reaching for it first. The founders started building years before ChatGPT, betting transformers would change how information is accessed. [2, 3]

Emerging Teams

- **Lab0:** Lab0 is building an AI forward deployment engineer for enterprise software, automating client process discovery, configuration, testing, and go-live. The key datapoint is implementation speed: YC says deployment cycles fall from six months to ten days. Founders: Onkar Borade, tokenaware, and Sujay Sriv. [4]

- **InLoopRobotics:** InLoopRobotics is selling warehouse automation as a monthly service rather than capex: packing, kitting, and fulfillment with no integrators and no 6-month PoC. Paid pilots are already live at 300+ picks per hour. Founders: FeduniakS, Zakariea_sh, and Pasha Rizali. [5]
- **Armature:** Armature is an early signal that "agent experience" may become its own software category. It runs real agent workflows to monitor and optimize how AI agents experience products, with a focus on improving MCP or CLI surfaces. Founders: Totzenberger and Louis Scremin. [6]
- **AI code-review tooling is starting to cluster:** YC-backed Stage is a guided code-review platform for understanding AI-generated code and claims faster review than GitHub, while Prix AI independently pitches AI as the first reviewer on GitHub PRs, flagging repetitive issues such as edge cases, logic mistakes, performance, security, and style problems before humans step in. The overlap suggests a real wedge is forming around QA for AI-written software. [7, 8]

AI & Tech Breakthroughs

- **Baseten's "owned intelligence" thesis is getting production proof points:** Baseten describes its stack as production-grade inference for companies moving from rented to owned intelligence by post-training models on their own application data. It cited Abridge, Decagon, OpenEvidence, Cursor, and Intercom as companies already adopting this pattern, and its technical work is pushing toward continual learning for long-horizon agentic tasks where models evolve with real-time data, tools, and specialized evals. [9]
- **Cohere Command A+:** Cohere said Command A+ is its most powerful LLM yet, optimized to run on minimal hardware and released as the company's first fully open-source Apache 2 model. For investors, it is a clean signal that efficient open models are still improving at the high end. [10, 11]
- **Context control is turning into real infrastructure:** Compresh reports roughly 60% fewer input tokens on long agent sessions by keeping the last four rounds raw and compressing older context into a partitioned memory view; in separate architecture writing, an "Adaptive Agent Architecture" proposes state-driven micro-agents, hard retry limits, and reflection anchors, with a claimed reduction from 15,000-50,000 tokens per task to 3,000-7,000. The broader takeaway is that memory and retry control are becoming first-class product surfaces. [12, 13]
- **Efficient-model research keeps moving:** A BitNet 1.58 writeup highlighted a ternary-weight approach using {-1, 0, +1} instead of FP16/FP32 weights, trading precision for higher dimensionality to preserve output

quality while reducing memory and compute demands. [14]

Market Signals

- **The competitive layer is moving above the base model:** Railway argues agent workloads need tighter control over network, compute, storage, orchestration, versioning, observability, and branching at 1,000x human scale; Armature is explicitly measuring how agents experience products; and a Reddit discussion around Google's enterprise agent platform framed the shift as moving from model hosting toward orchestration, governance, and multi-agent tooling. [15, 6, 16, 17, 18]

  "Pull request is definitely dying." [15]

- **Compute scarcity is creating infra moats:** Railway says its own bare-metal data centers deliver roughly three-month payback and ~70% margins, while cloud bursting across five providers helps avoid compute bottlenecks. Baseten says capacity constraints are worse than most outsiders think and has responded by distributing inference across 15-20 clouds and 80-100 regions. [15, 9]

- **Search and distribution are being rebuilt for AI agents:** Exa's financing rests on the idea that agent-first search wins hard, long-tail queries, while Georion is building a growth dashboard around AI visibility scanning, prompt tracking, AI crawler logs, and revenue attribution across engines such as ChatGPT, Claude, and Perplexity. [2, 3, 19]

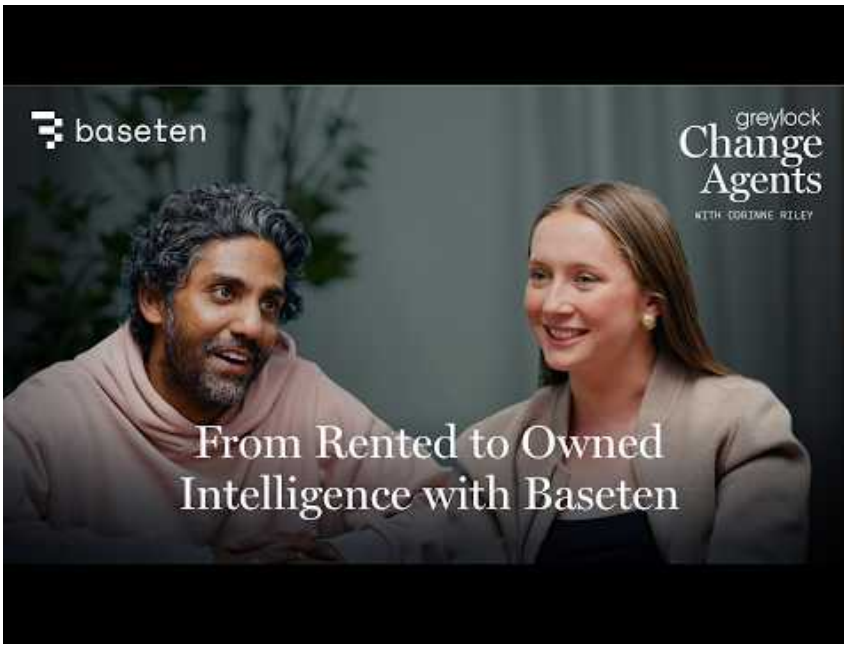
- **Capital structure may matter more than many app founders expect:** Gavin Baker argued that disaggregating prefill and inference could extend GPU useful lives from 3-4 years to 10-15 years, lowering financing rates and helping fund the AI buildout; in the same discussion, he said TSMC's capacity decisions are the key indicator for whether AI infrastructure turns into an overbuild. [20, 21]

- **Policy risk is rising around frontier releases:** Bindu Reddy flagged a planned White House executive order requiring frontier models to be reviewed 90 days before release, and argued it would boost China and open-source AI. [22]

Worth Your Time

- **Railway: The Agent-Native Cloud — Jake Cooper:** Best operator-level read in this batch on own-metal economics, cloud bursting, and what agents actually need from infrastructure. [15]

- **From Rented to Owned Intelligence with Baseten:** Best source here on post-training custom models, continual learning, and the move from



rented to owned intelligence. [9]
From Rented to Owned Intelligence with Baseten (8:00)

- **Gavin Baker on Orbital Compute, TSMC, and Frontier Models:** Strong macro/infra watch if you care about chip strategy, GPU financing, frontier-model economics, and bubble risk. [21]

Gavin Baker on Orbital Compute, TSMC, and Frontier Models (23:10)

- **GBrain thread** and **follow-up:** Quick read on open-source agent memory infrastructure, benchmarked long-memory performance, and context-engineering-driven idea generation. [23, 24]
- **Blank Bio seed announcement:** Short, useful read if you want the cleanest primary-source framing for the RNA-seq foundation-model thesis in clinical trials. [1]

Sources

1. X post by @ycombinator
2. X post by @ExaAILabs
3. X post by @a16z
4. X post by @ycombinator
5. X post by @ycombinator
6. X post by @ycombinator
7. X post by @ycombinator
8. r/SaaS post by u/theme-man
9. From Rented to Owned Intelligence with Baseten
10. X post by @cohere
11. X post by @aidangomez
12. r/SideProject post by u/talatt
13. r/SaaS post by u/Ok_Commission_8260
14. r/deeplearning post by u/Objective-Cash2188
15. Railway: The Agent-Native Cloud — Jake Cooper
16. r/artificial post by u/Few-Engineering-4135
17. r/artificial comment by u/SkyInfinite6282
18. r/artificial comment by u/Suspicious_Coat3244
19. r/SaaS post by u/Forward_Wind_8282
20. X post by @InvestLikeBest
21. Gavin Baker on Orbital Compute, TSMC, and Frontier Models
22. X post by @bindureddy
23. X post by @garrytan
24. X post by @garrytan