LeCun's AMI Labs, Agent Infrastructure, and the Turn Toward AI Efficiency

VC Tech Radar

2026-05-16

LeCun's AMI Labs, Agent Infrastructure, and the Turn Toward AI Efficiency

By VC Tech Radar • May 16, 2026

The strongest signals were Yann LeCun's new AMI Labs thesis around world models, a fresh wave of agent infrastructure and workflow startups, and growing evidence that smaller models, routing, and price-performance are reshaping AI architecture decisions. Investor appetite was most visible in healthcare AI and post-LLM technical bets.

1) Funding & Deals

- **AMI Labs:** Yann LeCun recently launched AMI, focused on world models and scaling the JEPA architecture he pioneered at Meta; he said investors were receptive because many were already recognizing LLM limitations and were interested in funding next-generation AI systems. The company is headquartered in Paris with a New York office. [1]
- **Healthcare AI investor demand:** Jay Rughani said he wants to fund healthcare applications that deliver more care and less paperwork, citing CounselHealth, EvidenceOpen, and Tennr as examples. Andrew Chen amplified the broader behavioral shift, arguing that *Dr AI* is already part of how people check decisions before and after doctor visits. [2, 3]
- **Angel-readiness signal:** A piano-teacher scheduling SaaS with 142 paying users and $1,280 MRR is now in conversations with two angel investors after the founder used Gamma to generate multiple deck framings and converged on the version an advisor said actually reflected the business. [4]

2) Emerging Teams

- **AMI Labs:** LeCun brings unusual pedigree for a seed-stage company: he shared the Turing Award and previously built FAIR before becoming Meta's chief AI scientist. His roadmap is to demonstrate hierarchical, action-conditioned world models across video and partner datasets within roughly 12-18 months, with target use cases spanning robotics, industrial process control, and healthcare. [1]
- **Montage:** The founder previously led product at Booking.com's restaurant-reservations platform across 100k+ venues. Montage's core product lets users edit video by editing transcript text, keeps footage at full 4K on files up to 20GB, and uses brief-based clip generation that one agency tester said cut review time from about three hours to about 40 minutes per video. [5]
- **AgentPhoneHQ:** YC-backed AgentPhoneHQ is positioning itself as telecom infrastructure for AI agents: one API gives each agent its own phone number and a trusted identity to reach the real world. YC highlighted founders @themeetmodi and @manav2modi at launch. [6]
- **Clicky:** YC also surfaced Clicky, a zero-setup consumer agent product from @FarzaTV that can see the user's screen, answer questions, make Notion docs, check Google Calendar, and create Linear tickets. [7, 8]
- **Lean operator signals:** Voremi's solo founder reports 500+ active users across India, the US, Pakistan, Brazil, Indonesia, the UK, and Nigeria with $0 ad spend for an AI voice reminder app built around fast voice input. RefundRadar is still early, but the workflow wedge is notable: it scores Shopify orders across 20+ risk signals before fulfillment and flags risky orders so merchants can hold or cancel before shipment. [9, 10, 11]

3) AI & Tech Breakthroughs

- **World models as the post-LLM thesis:** LeCun's core argument is that agentic systems need models that predict the consequences of actions and plan via search, not autoregressive next-token prediction. JEPA-style training predicts in abstract representation space rather than pixels, and he is targeting real-world control problems where action-conditioned models can optimize complex systems. [1]
- **30B-A3B reasoning model:** A newly released 30B-A3B model reportedly reached gold-medal level on IPhO and on IMO/USAMO with test-time self-verification and refinement, using what its authors describe as a simple unified scaling recipe for proof search. [12]
- **Efficient pre-training:** Nous Research published *Efficient Pre-Training with Token Superposition*, adding another concrete efficiency-oriented research thread to watch. [13]

- **Agentic software development at scale:** OpenClaw described a development stack that runs roughly 100 Codex instances for PR and issue review, security checks, issue deduplication, benchmark regression reporting, meeting-driven feature work, and auto-generated PRs when new issues fit the documented product vision. The team says the automation allows it to run the project extremely lean. [14]

4) Market Signals

- **Efficiency is becoming a first-class architecture theme:** A detailed essay in the batch argued that the frontier-only story is more a financing narrative than a production architecture one, pointing to $112B of Q1 2026 hyperscaler capex and $650-725B in full-year guidance on one side, but Phi-4, RouteLLM, and AWS Bedrock routing savings on the other. The same piece claims 40-60% of production token budgets are wasted by defaulting to frontier models, while 37% of enterprises with production AI already run five or more models. [15]
- **China's open-source pressure is increasingly about price-performance:** Bindu Reddy argued that Chinese pragmatic open-source models already handle 50% of everyday tasks at 30x lower cost, could handle most professional tasks within months, and present a catch-up challenge for US players. [16]
- **Healthcare is showing both user pull and investor pull:** Andrew Chen argued that consumers now routinely consult LLMs before and after seeing doctors, while Jay Rughani is actively looking to fund software that increases care delivery and reduces paperwork. [3, 2]
- **AI tools are compressing founder execution loops:** One non-technical founder said ChatGPT, tutorials, documentation, and trial-and-error were enough to build a working YouTube workflow SaaS in about three months. Separately, an AI pitch-deck generator let another founder test three financing narratives in one afternoon and choose the only framing that matched the actual business. [17, 4]

5) Worth Your Time

- **Yann LeCun on What Comes After LLMs** — useful for investors tracking world models, JEPA, and real-world AI applications in robotics

and industrial control. [1]
Yann LeCun on What Comes After LLMs (25:23)

- **OpenClaw's development thread** — useful for devtools investors because it lays out a stack of roughly 100 Codex instances across review, security, benchmarking, issue handling, and meeting-driven PR creation, with the team saying the automation lets it run lean. Marc Andreessen amplified the thread. [14, 18]

  How would we build software in the future if tokens don't matter? [14]

- **The Frontier-Only Narrative Is a Financing Story, Not an Architecture Story** — useful as a compact argument for smaller models, routing, and multi-model production stacks. [15]
- **30B-A3B reasoning model paper** — relevant if you track test-time self-verification, refinement, and proof-search scaling. [12]

Sources

1. Yann LeCun on What Comes After LLMs
2. X post by @JayRughani
3. X post by @andrewchen
4. r/SaaS post by u/Lopsided_Touch_4084
5. r/SideProject post by u/x_philomath_x
6. X post by @ycombinator

7. X post by @ycombinator
8. X post by @ycombinator
9. r/SaaS post by u/Much_Pomegranate6272
10. r/SaaS post by u/Electrical-Bus5079
11. r/SaaS comment by u/monishkurrra
12. X post by @stingning
13. r/deeplearning post by u/RecmacfonD
14. X post by @steipete
15. r/artificial post by u/gastao_s_s
16. X post by @bindureddy
17. r/SaaS post by u/ThinkingLoud99_
18. X post by @pmarca