

Physical AI Moves From Demo to Deployment as Open Models Compress the Stack

VC Tech Radar

2026-08-04

Physical AI Moves From Demo to Deployment as Open Models Compress the Stack

By VC Tech Radar • August 4, 2026

Andromeda’s clinical-autonomy round and Waymo’s operating evidence anchor a period in which safety, inference and power access matter as much as model quality. DeepSeek’s post-training leap and an AI-worm proof of concept sharpen both the upside and the diligence burden.

1. Funding & Deals

Andromeda Surgical raised a \$15M Series A led by Standard Capital, with Y Combinator and Vox Capital participating, taking total funding to \$30M. The company says its autonomous-surgery platform has treated 44 patients, lets the surgeon set the plan while the platform executes, and is cleared to launch in New Zealand and Canada; the round funds initial launch and scale-up. [1]

Standard Capital’s Dalton Caldwell says he has known founder Nick Damian since funding his first startup, Zenflow, and considers Damian the most knowledgeable founder in the YC network on healthcare regulatory processes. That background is unusually relevant here: the next diligence milestone is clinical and regulatory execution, not another autonomy demo. [2]

20VC disclosed a \$10M investment in Fireworks AI, a serving and fine-tuning bet rather than another foundation-model lab. The public investment memo says Fireworks was founded in 2022 by Li Qiao and six co-founders, largely the team that built and ran PyTorch at Meta; it reports 400-plus open models on the platform, kernel-level GPU optimization, and fine-tuning on customer data. The memo also reports more than \$1B in run-rate revenue and a customer moving a flagship feature from GPT-4o to an open model at 70–80% lower cost. Those traction figures are investor-reported and

should be verified, but the deal illustrates capital moving toward the layer that makes rapid model churn usable in production. [3, 4]

2. Emerging Teams

LangChain is productizing the operational layer around agents. Its managed DeepAgents offering is moving to public beta with Harbor-based evaluations, agent- and user-level memory, OAuth for tool access, Slack and GitHub integrations, and sandbox integration. The wedge is deliberately unglamorous: own the reliability and deployment plumbing so customers can concentrate on agent logic. [5]

Radiant Nuclear is worth tracking as a power-constrained infrastructure bet, with a material source caveat. The author of its current piece explicitly says he works at Radiant and is biased toward nuclear. The company says its 1-MW Kaleidos microreactor fits in a shipping container, has fuel at Idaho National Laboratory for full-power testing as the first new US reactor design in the DOME test bed, and is targeting a factory capable of 50 units per year with customer deliveries in 2028. The investment question is whether manufacturing and regulatory execution can make the factory-built thesis real. [6]

3. AI & Tech Breakthroughs

DeepSeek Flash’s latest update is a post-training signal, not yet a settled frontier-model claim. The Two Minute Papers transcript says the update arrived about three months after the prior system, with many benchmark results more than doubling and one improving sevenfold. It attributes the gain to post-training rather than a larger architecture: the underlying model size and architecture were unchanged, and the updated Flash reportedly beat a Pro model about five times larger. [7] The transcript also describes downloadable, permanently ownable weights with no usage caps and costs far below frontier providers. [7] The counter-signal is important: Abacus AI CEO Bindu Reddy calls Flash non-frontier, worse than Grok 4.5 in practice, and strong mainly at benchmark maximization. [8] The investable takeaway is to diligence post-training and task-specific evaluations separately from parameter count or headline benchmark gains.

A proof-of-concept AI worm makes local inference an infrastructure-security issue. Import AI reports work by researchers from the University of Toronto, Vector Institute, Cambridge, and ServiceNow on a worm that uses compromised GPU resources to host an open-weight LLM, detect vulnerabilities, tailor attacks, and infect additional hosts without relying on a vendor API. It fits on a single local A100 and reportedly achieved roughly 80% vulnerability detection, 53% exploitation, 88% self-replication, and about 37% full-attack success. [9] The important diligence detail is not just autonomous exploitation;

it is the combination of open weights, stolen compute, and independence from a provider that could revoke or monitor access.

Current research evidence splits AI’s verifiable capability from its open-ended judgment. Import AI reports that an internal version of OpenAI’s Astra solved ten open problems across mathematics and theoretical computer science. [9] In a separate shadow evaluation, Claude Opus 4.8 agents could perform the engineering needed for two unpublished research projects, but the original authors rejected both papers for poorly motivated experiments, no novel contribution, and impenetrable prose. [9] For recursive-research claims, independent novelty and expert acceptance remain more informative than successful task completion alone.

4. Market Signals

Inference engineering is becoming the control plane for model churn. Latent Space says the discipline barely existed as a category three years ago and now focuses on turning trained weights into products that are fast, reliable, and affordable at scale. It cites a GLM-5.2 experiment in which more quantization preserved benchmark quality while raising throughput 20%, and says optimization gains of 20%, 100%, or even 200% remain possible. [10] The same discussion describes grafting Kimi’s vision encoder onto GLM-5.2 without changing the base language model, [10] and a loop in which GLM-5.2 profiled and wrote GPU kernels for its own inference engine. [10] This is a durable infrastructure opportunity even if individual model weights commoditize.

Open versus closed models is becoming an enterprise procurement and sovereignty question. A current discussion frames AI sovereignty as owning the full supply chain, fine-tuning open models on company data, and avoiding dependence on a provider that may later compete with the customer. [11] The counter-signal from investor @altcap is that open and frontier models are not necessarily zero-sum: both may be needed to finance the projected US AI build-out. [12] Underwrite model portability, data rights, and geopolitical exposure rather than assuming that “open” automatically means lower total risk.

Physical AI’s moat is accumulated evidence, not the demo. Waymo’s transcript says its demo took 18 months but the product took about 15 years; it now reports more than 20 million autonomous trips, more than 200 million autonomous miles, and a scale of roughly half a million trips per week. [13] The company identifies four structural gaps versus digital AI—cost of error, latency, data, and validation—and says a car can travel about 100 feet in one second, forcing inference on board. [13] Waymo further argues that its safety framework and hundreds of millions of miles of publicly supported evidence are harder to replicate than its models or algorithms, reporting roughly 17 times fewer serious-injury crashes than human drivers over 220 million miles. [13] For early physical-AI companies, deployment evidence and validation infrastructure

belong in the moat analysis from day one.

Power access is becoming a compute-underwriting variable. Radiant’s article says utility forecasts for new US peak power over the next five years rose from 24 GW in 2022 to 166 GW in 2025, driven mostly by data centers, while acknowledging that some projects are double-counted. [6] The interconnection queue contains more proposed capacity than the country has built, only 13% of projects seeking connection from 2000–2020 reached operation, waits now run about five years, and high-voltage transmission construction fell from roughly 4,000 miles in 2013 to 55 miles in 2023. [6] The data argues for treating power contracts, interconnection and transmission timelines as core operating assumptions—not infrastructure footnotes.

Agent startups are finding that accountable execution beats maximal autonomy. An AI-sales builder says producing more output was easy; value shifted to account signals, evidence, and approval controls. [14] A separate builder found every customer wanted to preview an outbound message before sending it, while an AI-ops operator added named checkpoints for anything destructive or externally visible. [15, 16] Evidence, reversibility, and approval are becoming product features rather than concessions to low model quality.

Public markets are sorting AI beneficiaries rather than reopening the B2B software window. SaaStr’s current analysis reports that no venture-backed B2B unicorn had filed to go public in 2026, while arguing that consumption-priced infrastructure AI consumes more of is winning and seat-priced application software AI might replace is losing. [17] That is a useful exit-market filter for private software underwriting.

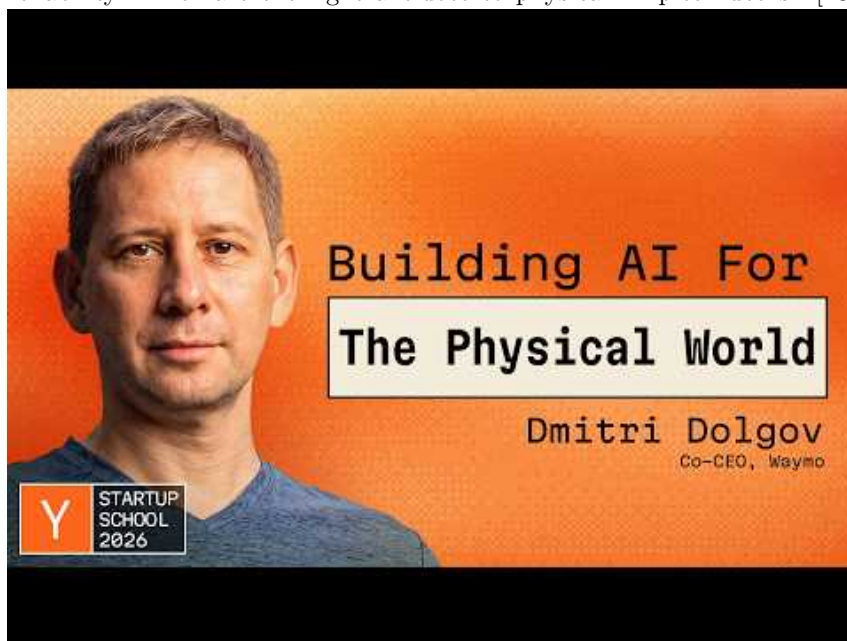
5. Worth Your Time

- **Watch *Another DeepSeek Moment Has Arrived*.** The useful segment is the explanation of how post-training changes behavior without changing the base model; read it alongside the counter-signal above. [7]



Another DeepSeek Moment Has Arrived (0:03)

- **Watch Waymo co-CEO Dmitri Dolgov's Startup School talk.** The demo-to-product gap and the exponential cost of each additional reliability "nine" are the right antidote to physical-AI pitch decks. [13]



Waymo Co-CEO Dmitri Dolgov: The Demo Is Only 1% Of The Work

(9:50)

- **Read/listen to The Inference Engineering Masterclass.** It is a practical map of quantization, model composition, KV-cache movement, and the infrastructure work between an open-weight release and a production API. [10]
 - **Read Import AI 467.** The value is the juxtaposition of an adaptive AI-worm capability, frontier-lab calls for pacing tools, and evidence that agents remain weak at open-ended research. [9]
-

Sources

1. X post by @andromedasurg
2. X post by @daltonc
3. X post by @HarryStebbing
4. X post by @P_Bonnet
5. X post by @hwchase17
6. X article by @torishiv
7. Another DeepSeek Moment Has Arrived
8. X post by @bindureddy
9. Import AI 467: Self-sustaining AI viruses; pacing AI progress; confusion about AI and creativity
10. The Inference Engineering Masterclass — Philip Kiely & Ali Taha, Baseten
11. X post by @HarryStebbing
12. X post by @altcap
13. Waymo Co-CEO Dmitri Dolgov: The Demo Is Only 1% Of The Work
14. r/EntrepreneurRideAlong post by u/Harshit-24
15. r/EntrepreneurRideAlong comment by u/akl773
16. r/EntrepreneurRideAlong comment by u/Positive-Emu-8379
17. 2026 Is The Biggest IPO Year In History. But Take Out SpaceX And Right Now It's One of the Worst.