MatX's $500M chip push, Meta's ~6GW AMD GPU buildout, and agents hardening into production

AI News Digest

2026-02-25

MatX's $500M chip push, Meta's ~6GW AMD GPU buildout, and agents hardening into production

By AI News Digest • February 25, 2026

A compute-and-agents day: MatX raises $500M for an LLM-specific chip while Meta commits ~6GW of AMD Instinct GPU capacity. In parallel, agent workflows harden (OpenAI WebSockets; Claude Code's measurable footprint) as safety/governance and IP tensions remain active (Anthropic RSP v3; Bengio's Law Zero; Marcus on unresolved training-data IP).

Today's threads to track

A clear throughline today: **AI's bottlenecks are moving down-stack** (memory, compute, silicon) at the same time that **agents are moving up-stack** (from coding into broader enterprise workflows). Several announcements—chips, data center buildouts, agent tooling, and safety policy—snap into that picture.

Compute & hardware: purpose-built LLM infrastructure accelerates

MatX raises $500M to build an LLM chip optimized for throughput *and* latency

MatX announced **MatX One**, an LLM chip it says targets **higher throughput than any announced system** while also matching **SRAM-first** latency and supporting **HBM** for long context, using a **splittable systolic array** plus a "fresh take on numerics" [1]. The company also disclosed a **$500M Series B** to finish development and scale manufacturing, with a **tapeout in under a year** [2].

Why it matters: This is a large, concrete bet that **LLM workloads are stable enough** to justify custom silicon—and that the "SRAM-first vs. HBM-first" tradeoff can be engineered around for long-context, agentic inference loops [3][4].

Meta signs multi-year deal to deploy AMD Instinct GPUs at ~6GW scale

Meta announced a multi-year agreement with AMD to integrate the latest **Instinct GPUs** into Meta's infrastructure, with ~**6GW of planned data center capacity** dedicated to the deployment [5].

Why it matters: The sheer capacity number is a signal of continued hyperscaler-scale buildout, reinforcing that **compute availability** remains a primary constraint on model development and deployment [6].

"Memory crowd-out" keeps surfacing as a practical limiter on agents and consumer tech

Ben Thompson argues AI is reviving a **thin-client** paradigm—chat and agent workflows that run in data centers, largely independent of local device capability [7][8]. He also frames an AI-driven **memory shortage** as a consumer-facing impact point as memory makers prioritize **HBM for AI chips**, pushing costs into broader electronics [9].

[1] post by @reinerpope
[2] post by @reinerpope
[3] post by @karpathy
[4] post by @reinerpope
[5] post by @AIatMeta
[6] post by @AIatMeta
[7] The Memory Crowd-Out | Stratechery by Ben Thompson
[8] The Memory Crowd-Out | Stratechery by Ben Thompson
[9] The Memory Crowd-Out | Stratechery by Ben Thompson

The Memory Crowd-Out | Stratechery by Ben Thompson (7:59)

Why it matters: If memory (HBM/DRAM/flash) is a gating factor for larger-context inference, it strengthens the gravitational pull toward **centralized data centers**—and can raise prices across non-AI hardware categories [10][11].

The "compute bottleneck" is being called out explicitly

Logan Kilpatrick said the compute bottleneck is "massively under appreciated," guessing the supply/demand gap is growing by a **single-digit percent every day**, and predicting it will **rate-limit AI's impact** on the economy and society [12][13].

Why it matters: This frames compute not just as a cost line item, but as the **macro constraint** determining how quickly agentic systems can spread into real workflows [14].

---

[10] The Memory Crowd-Out | Stratechery by Ben Thompson
[11] The Memory Crowd-Out | Stratechery by Ben Thompson
[12] post by @OfficialLoganK
[13] post by @OfficialLoganK
[14] post by @OfficialLoganK

Agents in production: from coding to “units of labor” across industries

Jack Clark: agents are shifting from “talkers” to “doers,” with multi-agent coordination becoming normal

In a discussion of AI agents’ economic impact, Jack Clark described a product arc from 2023–2024 “talkers” to 2026–2027 “doers” that can work together and oversee each other [15]. He also gave examples of internal productivity patterns—multiple “Claudes” reading documentation, summarizing it, and helping two people execute what would previously have required more time and coordination [16].

‘This is the pivotal point.’
How can I help you this evening?

How Fast Will A.I. Agents Rip Through the Economy? | The Ezra Klein Show (1:17)

Why it matters: This is a crisp articulation of the **agent product thesis**: workflows where the user specifies a goal and orchestration happens largely out of view—raising the value of instrumentation, oversight, and safety controls as autonomy grows [17][18].

[15] How Fast Will A.I. Agents Rip Through the Economy? | The Ezra Klein Show

[16] How Fast Will A.I. Agents Rip Through the Economy? | The Ezra Klein Show

[17] How Fast Will A.I. Agents Rip Through the Economy? | The Ezra Klein Show

[18] How Fast Will A.I. Agents Rip Through the Economy? | The Ezra Klein Show

OpenAI adds WebSockets to the Responses API for long-running, tool-heavy agents

OpenAI introduced **WebSockets in the Responses API**, positioned for “low-latency, long-running agents with heavy tool calls” [19]. Greg Brockman said it yields **30% faster agentic rollouts in Codex** [20].

Docs: http://developers.openai.com/api/docs/guides/websocket-mode [21]

Why it matters: This is infrastructure aimed directly at **agent runtime performance**, suggesting that “agent UX” improvements increasingly come from **systems plumbing**, not just model quality [22][23].

Claude Code’s one-year mark: measurable footprint + new “remote control” workflow

A Latent Space / SemiAnalysis discussion says Claude Code (launched Feb 24, 2025) is now responsible for ~**4% of GitHub code** [24][25]. Separately, a new **/remote-control** feature lets users continue local Claude Code sessions from a phone, rolled out to Max users [26].

Why it matters: “Share of GitHub” is an early, imperfect—but concrete—signal that coding agents are moving from demos into routine practice, and that labs are investing in **always-available, multi-device agent workflows** [27][28].

Devin (Cognition) focuses on enterprise-proven UX and “closing the loop”

Swyx reported that **Devin 2.2** is a self-serve UX overhaul, integrating an omnibox and tying “Devin Review” back into the main agent to “close the loop” [29]. He also shared enterprise usage growth claims: per-enterprise usage doubled every 2 months in 2025, accelerating to every 6 weeks this year, with internal usage at 4× the 2025 peak [30].

Why it matters: Even if individual metrics are anecdotal, the emphasis is notable: agent products competing on **workflow design + iteration loops**, not just

[19] post by @OpenAIDevs

[20] post by @gdb

[21] post by @OpenAIDevs

[22] post by @OpenAIDevs

[23] post by @gdb

[24] Claude Code for Finance + The Global Memory Shortage: Doug O’Laughlin, SemiAnalysis

[25] Claude Code for Finance + The Global Memory Shortage: Doug O’Laughlin, SemiAnalysis

[26] post by @_catwu

[27] Claude Code for Finance + The Global Memory Shortage: Doug O’Laughlin, SemiAnalysis

[28] post by @_catwu

[29] post by @swyx

[30] post by @swyx

raw coding ability [31].

"Build for agents": Karpathy spotlights CLIs and agent-accessible surfaces

Karpathy amplified the idea that "legacy" interfaces like **CLIs** are attractive because agents can use them directly—installing tools, composing terminal utilities, and building dashboards quickly [32]. He also urged product builders to ensure docs are exportable (e.g., markdown) and that services are usable via CLI or MCP: "It's 2026. Build. For. Agents." [33].

Why it matters: This is a practical distribution lesson: products that expose **agent-friendly primitives** (CLI/APIs/skills) are easier to integrate into emerging agent ecosystems [34].

Accounting joins the long-horizon agent wave: Basis raises $100M at $1.15B

Basis (trybasis) said it raised **$100M at a $1.15B valuation** to deploy accounting agents across CAS, tax, audit, and advisory [35]. The company claims adoption by **30% of the Top 25 accounting firms** and reported an "accounting agent" completing a **business tax workbook end-to-end** [36][37].

Why it matters: This is a milestone claim for **non-coding, regulated knowledge work** being tackled with "production-grade, long-horizon agents" [38][39].

Safety, governance, and the geopolitics/IP backdrop

Anthropic updates its Responsible Scaling Policy to v3 and commits to more transparency artifacts

Anthropic announced **Responsible Scaling Policy (RSP) v3**, saying it incorporates lessons since 2023 and commits to "even greater transparency" [40]. The update includes publishing **Frontier Safety Roadmaps** (detailed safety goals) and **Risk Reports** that quantify risk across deployed models, and separating unilateral commitments from industry recommendations [41][42].

[31] post by @swyx
[32] post by @karpathy
[33] post by @karpathy
[34] post by @karpathy
[35] post by @trybasis
[36] post by @trybasis
[37] post by @trybasis
[38] post by @trybasis
[39] post by @trybasis
[40] post by @AnthropicAI
[41] post by @AnthropicAI
[42] post by @AnthropicAI

Announcement: https://anthropic.com/news/responsible-scaling-policy-v3 [43]

Why it matters: This continues a shift toward **published, structured safety commitments** that can be compared over time—moving beyond one-off statements into repeatable governance outputs [44].

Bengio's "Law Zero": safe-by-design AI as a distinct R&D track

Yoshua Bengio described founding **Law Zero**, a nonprofit AI lab with **>$30M philanthropic funding**, focused on designing AI systems that "will not harm people" and exploring ways to disentangle "world understanding" from agency/intentions [45][46]. He also argued for transparency-based regulation (citing the EU as leading) and emphasized international coordination and incentives like insurance [47][48].

Why it matters: This is an attempt to build **institutional capacity** around safety-first architectures and policy mechanisms, rather than treating safety purely as a constraints layer on frontier labs [49][50].

IP tensions remain unresolved even as "model protection" becomes a national-security talking point

Gary Marcus argued the foundation model industry sits on an unresolved IP question, noting Anthropic settled **$1.5B** over **7M pirated books** and claiming "every lab trained on data it did not license" [51][52]. He also pointed to the irony of US export controls framed around IP while domestic model training practices remain contested [53].

> "watching billionaires argu[ing] about who stole … more ethically" [54]

Why it matters: As labs tighten access and frame capability protection geopolitically, the domestic IP foundation remains a live vulnerability—legally and rhetorically [55][56].

---

[43] post by @AnthropicAI

[44] post by @AnthropicAI

[45] Yoshua Bengio - Fireside Chat with Yoshua Bengio [Alignment Workshop]

[46] Yoshua Bengio - Fireside Chat with Yoshua Bengio [Alignment Workshop]

[47] EU Leading on AI Regulation, Says Canadian Computer Scientist Yoshua Bengio | Global AI Lens

[48] Yoshua Bengio - Fireside Chat with Yoshua Bengio [Alignment Workshop]

[49] Yoshua Bengio - Fireside Chat with Yoshua Bengio [Alignment Workshop]

[50] Yoshua Bengio - Fireside Chat with Yoshua Bengio [Alignment Workshop]

[51] post by @shanaka86

[52] post by @shanaka86

[53] post by @shanaka86

[54] post by @GaryMarcus

[55] post by @shanaka86

[56] post by @shanaka86

Research & model releases worth noting

Inception Labs ships "Mercury 2," described as a reasoning diffusion LLM

A post announcing Mercury 2 calls it the "world's first reasoning diffusion LLM," claiming **5× faster performance** than leading speed-optimized LLMs [57]. Andrew Ng called diffusion LLMs a "fascinating alternative" to autoregressive models and praised the inference speed [58].

Why it matters: If performance claims hold up in broader use, this is a notable productization step for **non-autoregressive** LLM families aimed at real-world latency constraints [59].

NVIDIA open-sources "SONIC" whole-body humanoid control (42M transformer)

NVIDIA's GEAR lab released **SONIC**, a 42M-parameter transformer for humanoid whole-body control, trained at scale (100M+ mocap frames; 500k+ parallel robots) and reported **zero-shot transfer** to a real robot with **100% success** across 50 motion sequences [60]. The project is released with paper/code/site [61][62][63].

Why it matters: This is a concrete, open-source datapoint for scaling simulation + imitation/RL pipelines into robust real-world humanoid motion control [64].

Open models: Qwen 3.5 adds both MoE and dense options

Three Qwen 3.5 models were highlighted: **122B-A10B (MoE)**, **35B-A3B (MoE)**, and a **27B dense** model [65]. Nathan Lambert argued dense releases are important for the open ecosystem until fine-tuning MoEs to a single domain is more broadly "distributed" [66].

Why it matters: This reflects ongoing experimentation in open-weight model form factors—balancing efficiency (MoE) with fine-tuning practicality (dense) [67].

[57] post by @StefanoErmon
[58] post by @AndrewYNg
[59] post by @StefanoErmon
[60] post by @DrJimFan
[61] post by @DrJimFan
[62] post by @DrJimFan
[63] post by @DrJimFan
[64] post by @DrJimFan
[65] post by @TheAhmadOsman
[66] post by @natolambert
[67] post by @natolambert

Fine-tuning data selection: targeted instruction selection framework (LESS + selectors)

A new preprint on targeted instruction selection separates (1) **representations** (e.g., gradient-based **LESS**) from (2) **selectors** (e.g., greedy round-robin, optimal transport), reporting that LESS distance correlates strongly with performance and offering a practical recipe by budget size [68][69][70].

Paper: https://arxiv.org/abs/2602.14696 [71] Code: https://github.com/dcml-lab/targeted-instruction-selection [72]

Why it matters: As more teams fine-tune task-specific models, systematic selection methods can be a lever for **quality per labeling/token dollar** [73].

Enterprise & public-sector deployment signals

Microsoft expands Sovereign Cloud for fully disconnected AI deployments

Microsoft announced new Sovereign Cloud capabilities that let customers bring **productivity workloads and AI models into fully disconnected sovereign environments**, emphasizing more local control and regulatory/security needs [74][75].

Details: https://blogs.microsoft.com/blog/2026/02/24/microsoft-sovereign-cloud-adds-governance-productivity- [76]

Why it matters: This targets a growing deployment constraint: customers who want frontier capabilities but require **sovereignty and isolation** by design [77].

NVIDIA healthcare survey: adoption rising; agentic AI enters the workload mix

NVIDIA's "State of AI in Healthcare and Life Sciences" survey reports **70%** of organizations actively using AI (up from 63% in 2024) [78] and **69%** using generative AI/LLMs (up from 54%) [79]. It also reports **47%** are using or assessing

[68] r/MachineLearning post by u/nihalnayak

[69] r/MachineLearning post by u/nihalnayak

[70] r/MachineLearning post by u/nihalnayak

[71] r/MachineLearning post by u/nihalnayak

[72] r/MachineLearning post by u/nihalnayak

[73] r/MachineLearning post by u/nihalnayak

[74] post by @satyanadella

[75] post by @satyanadella

[76] post by @satyanadella

[77] post by @satyanadella

[78] From Radiology to Drug Discovery, Survey Reveals AI Is Delivering Clear Return on Investment in Healthcare

[79] From Radiology to Drug Discovery, Survey Reveals AI Is Delivering Clear Return on Investment in Healthcare

agentic AI [80], while **85%** of executives say AI helps increase revenue and **80%** say it helps reduce costs [81].

Why it matters: This suggests the industry is moving from experimentation to execution—and that "agents" are now a named category being tracked in enterprise adoption data [82][83].

Quick hits

- **Perplexity Comet**: an upgraded voice mode is rolling out, described as enabling **fully hands-free browser control**, built with OpenAI's "latest real time model" [84][85].
- **OpenAI**: named **Arvind KC** as Chief People Officer, framing the hire around guiding AI-enabled work responsibly [86][87].
- **Google DeepMind**: launched a Europe-focused **Robotics Accelerator** with technical deep dives, mentorship, and up to **$350k** in Cloud credits [88][89].

Sources

1. post by @reinerpope
2. post by @karpathy
3. post by @AIatMeta
4. The Memory Crowd-Out | Stratechery by Ben Thompson
5. post by @OfficialLoganK
6. post by @OfficialLoganK
7. How Fast Will A.I. Agents Rip Through the Economy? | The Ezra Klein Show
8. post by @OpenAIDevs
9. post by @gdb
10. Claude Code for Finance + The Global Memory Shortage: Doug O'Laughlin, SemiAnalysis

[80] From Radiology to Drug Discovery, Survey Reveals AI Is Delivering Clear Return on Investment in Healthcare

[81] From Radiology to Drug Discovery, Survey Reveals AI Is Delivering Clear Return on Investment in Healthcare

[82] From Radiology to Drug Discovery, Survey Reveals AI Is Delivering Clear Return on Investment in Healthcare

[83] From Radiology to Drug Discovery, Survey Reveals AI Is Delivering Clear Return on Investment in Healthcare

[84] post by @AravSrinivas

[85] post by @AravSrinivas

[86] post by @OpenAI

[87] post by @OpenAI

[88] post by @GoogleDeepMind

[89] post by @GoogleDeepMind

11. post by @_catwu
12. post by @swyx
13. post by @karpathy
14. post by @trybasis
15. post by @AnthropicAI
16. post by @AnthropicAI
17. post by @AnthropicAI
18. Yoshua Bengio - Fireside Chat with Yoshua Bengio [Alignment Workshop]
19. EU Leading on AI Regulation, Says Canadian Computer Scientist Yoshua Bengio | Global AI Lens
20. post by @shanaka86
21. post by @GaryMarcus
22. post by @StefanoErmon
23. post by @AndrewYNg
24. post by @DrJimFan
25. post by @DrJimFan
26. post by @TheAhmadOsman
27. post by @natolambert
28. r/MachineLearning post by u/nihalnayak
29. post by @satyanadella
30. From Radiology to Drug Discovery, Survey Reveals AI Is Delivering Clear Return on Investment in Healthcare
31. post by @AravSrinivas
32. post by @AravSrinivas
33. post by @OpenAI
34. post by @GoogleDeepMind
35. post by @GoogleDeepMind