Agent Workflows Become Strategy, Policy, and Infrastructure

AI News Digest

2026-03-09

Agent Workflows Become Strategy, Policy, and Infrastructure

By AI News Digest • March 9, 2026

AI agents were the clear theme today: leading companies are turning them into workflow harnesses, research loops, and inference infrastructure, while Shenzhen is drafting policy around AI-native solo startups. Also notable, Microsoft showed a glass-based archival system that uses AI to read back data with zero errors.

Agents were the dominant story today

OpenAI and Cognition show how agent performance is becoming a harness problem

A small OpenAI team says it used Codex to open and merge 1,500 pull requests with zero manual coding to ship an internal product used by hundreds of internal users. swyx groups that with OpenAI's Frontier, Symphony, and harness engineering efforts as part of the emerging AI-native organization; in parallel, he says Cognition's Devin evaluates dozens of model groups and regularly rewrites its harness, while one user says Devin 2.2 now feels simpler for them to use basically all the time, even when a change starts locally [1][2][3][4].

> "Build a company that benefits from the models getting better and better" [5]

Why it matters: The edge is starting to shift from any single model to the evals, routing, and workflow systems wrapped around improving models. A

---

[1] post by @OpenAIDevs
[2] post by @swyx
[3] post by @swyx
[4] post by @dtcb
[5] post by @swyx

useful check on the narrative: Martin Casado says AI still struggles with finicky renderer work in sparkjs, where the main developer went back to hand-coding the renderer while keeping AI for tests, demos, and prototypes [6].

Karpathy is pushing autoresearch from a solo loop toward a research community

Karpathy says improvements found across roughly 650 experiments over two days on a depth-12 model transferred to depth-24, setting up a new nanochat leaderboard entry for "time to GPT-2." He also says the next step for autoresearch is asynchronous, massively collaborative agents—closer to a research community than a single PhD student—with GitHub Discussions and PRs as lightweight coordination surfaces [7891011].

Why it matters: This is a concrete extension of autonomous research: not just an agent editing training code, but many agents contributing branches, reading prior results, and feeding findings back into a shared repo. Repo: autoresearch [12]

Policy and infrastructure are starting to reorganize around agents

Shenzhen is drafting public support for AI-native "one person companies"

Longgang District in Shenzhen released a draft policy to support OpenClaw and the OPC model, where one person uses AI agents across R&D, production, operations, and marketing. The package includes public datasets, data-service subsidies, procurement support for OpenClaw-based solutions, free compute, subsidized workspace, relocation support, competition awards, and seed-stage equity investment up to RMB 10 million; the consultation window runs from March 7 to April 6, 2026 [131415161718192021].

---

[6] post by @martin_casado
[7] post by @karpathy
[8] post by @karpathy
[9] post by @karpathy
[10] post by @karpathy
[11] post by @karpathy
[12] post by @karpathy
[13] r/LocalLLM post by u/Alert_Efficiency_627
[14] r/LocalLLM post by u/Alert_Efficiency_627
[15] r/LocalLLM post by u/Alert_Efficiency_627
[16] r/LocalLLM post by u/Alert_Efficiency_627
[17] r/LocalLLM post by u/Alert_Efficiency_627
[18] r/LocalLLM post by u/Alert_Efficiency_627
[19] r/LocalLLM post by u/Alert_Efficiency_627
[20] r/LocalLLM post by u/Alert_Efficiency_627
[21] r/LocalLLM post by u/Alert_Efficiency_627

- Up to **RMB 10 million** in equity support for seed-stage OPC startups [22]
- **Three months of free compute** and project funding up to **RMB 4 million** for strong demonstration projects [23]
- Public datasets plus subsidies for data services and OpenClaw deployments [24][25]

Why it matters: This draft directly funds solo AI-agent startups rather than only general AI R&D. That makes it a notable economic-development signal around how local governments think the agent ecosystem may evolve [26][27].

Nvidia is treating agent inference as a systems problem

On Latent Space, Nvidia engineers described Dynamo as a data-center-scale inference layer on top of vLLM, SGLang, and TensorRT-LLM that uses disaggregation to separate prefill and decode, then adjusts worker ratios as workloads change. They also connected agent workloads to more structured contexts and better cache behavior, and previewed GTC sessions on Dynamo and "the future of agents in production inference" [28][29][30][31][32][33].

Why it matters: If agents impose more repeatable structure than chatbots, infra teams get new levers for speed and cost. The same conversation also emphasized sandboxing and permission boundaries: Brev provides one-click GPU provisioning and an isolated place to run tools like OpenClaw, while the security rule of thumb was to give agents only two of three powers—file access, internet access, and code execution [34][35][36][37].

---

[22] r/LocalLLM post by u/Alert_Efficiency_627

[23] r/LocalLLM post by u/Alert_Efficiency_627

[24] r/LocalLLM post by u/Alert_Efficiency_627

[25] r/LocalLLM post by u/Alert_Efficiency_627

[26] r/LocalLLM post by u/Alert_Efficiency_627

[27] r/LocalLLM post by u/Alert_Efficiency_627

[28] NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"

[29] NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"

[30] NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"

[31] NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"

[32] post by @swyx

[33] post by @swyx

[34] NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"

[35] NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"

[36] NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"

[37] NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"

NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light" (55:04)

Also notable

Microsoft shows long-term glass storage with AI-based readout

Microsoft's Project Silica writes 5 TB into ordinary glass across 301 layers using ultrafast lasers, then reads it back with microscope imaging and an AI image-recognition model that the company says decodes the data with zero errors. The storage medium requires no power to preserve the data and is described as resistant to heat, water, radiation, and magnetic fields, with accelerated testing projecting more than 10,000 years of room-temperature life [38][39].

Why it matters: This is storage infrastructure rather than a new model, but it is a meaningful Microsoft + Nature result aimed at the energy cost of archival data. For long-lived cloud archives, it points to a very different tradeoff than magnetic tape [40].

---

[38] post by @rowancheung

[39] post by @rowancheung

[40] post by @rowancheung

Sources

1. post by @OpenAIDevs
2. post by @swyx
3. post by @swyx
4. post by @dtcb
5. post by @martin_casado
6. post by @karpathy
7. post by @karpathy
8. post by @karpathy
9. r/LocalLLM post by u/Alert_Efficiency_627
10. NVIDIA's AI Engineers: Brev, Dynamo and Agent Inference at Planetary Scale and "Speed of Light"
11. post by @swyx
12. post by @rowancheung
13. post by @rowancheung