Brain Emulation, Agentic Tooling, and New Infrastructure Across AI

AI High Signal Digest

2026-03-09

Brain Emulation, Agentic Tooling, and New Infrastructure Across AI

By AI High Signal Digest • March 9, 2026

This brief covers Eon Systems' fruit fly connectome simulation, the shift toward harnessed and terminal-native agent workflows, new synthetic data and skill infrastructure, hardware aimed at agentic inference, and policy signals from New York and Shenzhen.

Top Stories

Why it matters: The most consequential updates this cycle centered on training inputs, agent scaffolding, deployment hardware, and governance. [1, 2, 3, 4]

1) Eon Systems pushed a connectome-driven fruit fly into a simulated body

Eon said it took the FlyWire connectome of the fruit fly brain, applied a simple neuron model, and used it to control a MuJoCo physics-simulated body, closing the loop from neural activation to action. [5]

Observers said the simulated fly showed walking, grooming, and feeding-like behaviors without training data or gradient descent, and one post described the result as what may be the first whole-brain emulation controlling a body. [6, 7]

The significance is methodological: the system is being framed as modeling neural structure rather than learning behavior from examples. [6, 7]

A note of caution came from another expert, who argued the work is still far from a biophysically faithful fly-brain simulation because individual neurons are much more complex than this setup captures. [8]

2) Agentic coding is becoming a systems discipline

The new OpenDev paper argues the field is shifting from IDE plugins to terminal-native agents and lays out concrete reliability patterns, including workload-specialized model routing, separate planning and execution agents, lazy tool discovery, adaptive context compaction, cross-session memory, and strict safety controls. [2]

That direction is showing up in operations as well: OpenAI said a small team steering Codex opened and merged 1,500 pull requests with zero manual coding for a product used by hundreds of internal users. [9]

LangChain's new LangSmith Skills + CLI extends the same idea by letting coding agents debug traces, create datasets, and run experiments natively in the terminal. [10]

At the application layer, Devin's team says its system evaluates a couple dozen model groups for harness inclusion and rewrites its stack every few months, while one user said version 2.2 now feels simpler than local development for most work. [11, 12]

3) Synthetic data and reusable skills are being treated as first-class assets

Hugging Face released FinePhrase and a Synthetic Data Playbook after more than 90 experiments and 1T generated tokens, producing a 500B-token synthetic dataset and publishing the associated recipes and code. [1]

SkillNet complements that effort on the agent side: it organizes more than 200,000 AI skills inside a unified ontology with relationships such as similarity, composition, and dependency, and reports a 40% improvement in average rewards with 30% fewer execution steps across ALFWorld, WebShop, and ScienceWorld. [13]

Together, these releases suggest teams are increasingly productizing the inputs to intelligence, not just the final model. Resources: https://huggingface.co/spaces/HuggingFaceFW/finephrase and https://arxiv.org/abs/2603.04448 [1, 13]

4) SambaNova launched hardware aimed directly at agentic inference

SambaNova introduced the SN50 RDU, presenting it as a chip designed for the cost profile of agentic inference rather than conventional GPU-style serving. [3]

The architecture maps model graphs directly onto hardware data paths and adds agentic caching across large-capacity memory, HBM, and SRAM so multiple models can stay resident and switch in milliseconds. [3]

Reported performance claims versus NVIDIA Blackwell B200 were 5× faster inference, 3× higher throughput, and up to 8× lower TCO on large models,

with SambaRack SN50 scaling to 256 accelerators and support for up to 10T-parameter models and 10M-token contexts. [3]

SN40L is available now, while SN50 and SambaRack SN50 are expected in H2 2026. [3]

5) OpenAI's robotics leadership change made autonomy concerns concrete

Caitlin Kalinowski resigned from OpenAI over concerns about "lethal autonomy without human intervention." She had led the robotics division after joining from Meta in November. [4]

> "This was about principle, not people." [4]

The resignation lands as robotics builders are also publicly describing unusually fast progress: Brett Adcock said he has "never seen this much progress in robotics" and that his lab is seeing capabilities emerge that "we didn't even know were possible." [14]

Research & Innovation

Why it matters: This cycle's research was unusually concrete about when agents help, how they should plan, and how automated research systems may scale. [15, 16, 17, 18]

Multi-agent gains depend on task structure

A study across 180 configurations found multi-agent setups can improve performance by up to 81% on parallelizable tasks such as financial analysis, but degrade performance by up to 70% on sequential tasks such as Minecraft crafting. [15]

The paper also fits an equation that predicts the best architecture for a new task 87% of the time. PDF: https://arxiv.org/pdf/2512.08296 [15]

Structured planning continues to outperform greedy web agents

StructuredAgent introduces dynamic AND/OR trees plus structured memory so agents can backtrack, revise, and preserve alternative solutions during long web tasks. [16]

It reports 46.7% success on complex shopping tasks and interpretable hierarchical plans that make debugging and human intervention easier. Paper: https://arxiv.org/abs/2603.05294 [16]

Automated research stacks are opening up

Google DeepMind said it is open-sourcing part of its automated-research infrastructure for Gemini in the repo https://github.com/google-deepmind/simply,

describing it as more complex than the nanochat setup but closer to state-of-the-art LLM pre- and post-training. [17]

Karpathy also described the next step for autoresearch as asynchronously, massively collaborative agents, more like a research community than a single PhD student, with experiments summarized in GitHub Discussions or PRs that agents can later read and build on. [18]

Model and tooling design notes

Hugging Face redesigned `transformers` to make mixture-of-experts models first-class citizens, covering weight loading, expert routing backends, parallelism, and training optimizations. [19]

A separate argument from world-model research said symbolic world models that abstract away from pixels are especially important for agents, while also acknowledging that converting real-world signals into symbols remains unsolved. [20]

Products & Launches

Why it matters: New launches this cycle focused on making agents easier to run locally, inspect, and integrate into everyday workflows. [10, 21, 22, 23]

- **Codex:** Recent updates included GPT 5.4, Windows support, Fast mode, and new skills such as Playwright Interactive, Slides, and Spreadsheets, alongside Codex Security and Codex for OSS. Official site: https://openai.com/codex/ [24, 25]
- **LangSmith Skills + CLI:** LangChain released Skills + CLI so coding agents can debug traces, create datasets, and run experiments from the terminal. More: https://blog.langchain.com/langsmith-cli-skills/ [10, 26]
- **OpenClaw on Jetson:** NVIDIA Robotics published a tutorial for running a fully local, always-on assistant on Jetson with zero cloud APIs; vLLM said the setup can serve MoE models such as Nemotron 3 Nano 30B on Jetson AGX. Tutorial: https://www.jetson-ai-lab.com/tutorials/openclaw/ [21, 27]
- **FireRed-Image-Edit-1.1:** fal launched a new image-editing model with identity consistency across edits, multi-image reference blending, portrait makeup, text style reference, and photo restoration. Try it here: https://fal.ai/models/fal-ai/firered-image-edit-v1.1 [22, 28]
- **Hermes Agent:** Nous Research published docs for Hermes Agent at https://hermes-agent.nousresearch.com/docs; earlier this week the app rose from #41 to #21 on OpenRouter. [23, 29]

Industry Moves

Why it matters: The clearest business pattern this cycle was investment in the operating layer around models: harnesses, routing, infra, and distribution. [9,

11, 30, 31]

AI-native organizations are standardizing around harnesses

OpenAI's Harness Engineering post said a small team used Codex to open and merge 1,500 pull requests with zero manual coding for a product used by hundreds of internal users. [9]

Devin's reported setup follows a similar logic: it uses a couple dozen model groups, evaluates models extensively for harness inclusion, and rewrites the stack every few months; one frequent user said Devin 2.2 now feels simpler than local development for most tasks. [11, 12]

> "Build a company that benefits from the models getting better and better" [11]

Infrastructure competition is widening

NVIDIA acquired Brev.dev, whose founders said they started the company to build the best possible developer experience and had already been working closely with NVIDIA since August. [30]

Huawei, meanwhile, showcased the Atlas 950 SuperPoD with 8,192 cards and the Atlas 850E inference server; one estimate said the SuperPoD is roughly comparable to 8K H200s, with Q4 2026 delivery constrained by HBM and NPU chip bottlenecks. [32, 33]

On the demand side, Similarweb said Claude was the fastest-growing generative AI tool by website visits in February. [31]

Policy & Regulation

Why it matters: Policy signals are still early, but this cycle included both a concrete disclosure rule and direct public subsidies for agent deployment. [34, 35]

New York added a clear disclosure and consent requirement

New York will require disclosure when AI is used in advertising and prior consent for the commercial use of a deceased individual's name, voice, or image. [34]

Shenzhen is subsidizing agent deployment directly

Shenzhen rolled out free OpenClaw setup, three months of free computing power, a 50% subsidy on data services, and a 30% hardware subsidy. One observer said the scale and direct government involvement make the security implications of agents harder to ignore. [35, 36]

Quick Takes

Why it matters: These smaller items help track where capability, tooling, and evaluation practice are moving next. [37, 38, 39, 40]

- **Claude-assisted debugging:** A Zhihu writeup said Claude Opus 4.6 helped isolate a DeepEP race condition involving PyTorch deterministic mode, GPU streams, and NaN-filled buffers after roughly two days of intermittent runs. [37]
- **Small-model pressure:** One tester concluded Qwen 3.5-4B is about as good as GPT-4o in most benchmarked cases; another said its reasoning version was narrowly stronger on WildChat but more verbose, less knowledgeable, and more hallucination-prone. [38, 41]
- **OpenClaw benchmarking:** PinchBench launched to compare model performance on OpenClaw-style tasks. [42, 43]
- **Secure execution:** Monty, a minimal secure Python interpreter written in Rust for AI use cases, is now on GitHub at https://github.com/pydantic/monty. [39]
- **Kernel optimization:** A fused RMS Norm + NVFP4 quantization kernel written in CuTeDSL reported a consistent ~2.9× speedup over separate Triton kernels. [44, 45]
- **LLM eval rigor:** A forthcoming long-form post on applied statistics for LLM evals highlighted noise reduction, more confident conclusions, and faster experiments, with paper recommendations attached. [40]

Sources

1. X post by @lvwerra
2. X post by @omarsar0
3. X post by @TheTuringPost
4. X post by @TheRundownAI
5. X post by @michaelandregg
6. X post by @oh_that_hat
7. X post by @kimmonismus
8. X post by @fabianstelzer
9. X post by @OpenAIDevs
10. X post by @LangChain
11. X post by @swyx
12. X post by @dtcb
13. X post by @omarsar0
14. X post by @adcock_brett
15. X post by @burkov
16. X post by @omarsar0
17. X post by @crazydonkey200
18. X post by @karpathy

19. X post by @dl_weekly
20. X post by @sirbayes
21. X post by @NVIDIARobotics
22. X post by @fal
23. X post by @Teknium
24. X post by @reach_vb
25. X post by @thsottiaux
26. X post by @LangChain
27. X post by @vllm_project
28. X post by @fal
29. X post by @Teknium
30. X post by @NaderLikeLadder
31. X post by @Similarweb
32. X post by @tphuang
33. X post by @teortaxesTex
34. X post by @francoisfleuret
35. X post by @bigmagicdao
36. X post by @teortaxesTex
37. X post by @ZhihuFrontier
38. X post by @N8Programs
39. X post by @dl_weekly
40. X post by @cwolferesearch
41. X post by @teortaxesTex
42. X post by @steipete
43. X post by @TheZachMueller
44. X post by @maharshii
45. X post by @maharshii