MiniMax M3 Launches as OpenAI Formalizes Robotics and NVIDIA Expands Open Models

AI High Signal Digest

2026-06-01

MiniMax M3 Launches as OpenAI Formalizes Robotics and NVIDIA Expands Open Models

By AI High Signal Digest • June 1, 2026

MiniMax's M3 dominated the day with an unusually broad open-weight release, while OpenAI formalized its robotics effort and NVIDIA expanded both open models and physical-AI infrastructure. The brief also covers standout research on training efficiency and context management, plus notable new multimodal products.

Top Stories

Why it matters: open models, robotics, and physical AI all moved closer to deployment today.

- **MiniMax M3 combined frontier coding/agent performance, 1M context, and native multimodality in one open-weight model.** MiniMax introduced M3 as the first open-weights model to combine those three capabilities, with benchmarks including 59.0% on SWE-Bench Pro and 66.0% on Terminal Bench 2.1 [1]. It was distributed quickly across OpenRouter, Ollama Cloud, and Together-powered inference, while weights and a technical report are due in about 10 days [2, 3, 4, 1].
- **OpenAI turned its world-simulation work into a robotics division.** OpenAI said its program led by Aditya Ramesh has evolved into OpenAI Robotics, which is hiring full-stack hardware, ops, systems, and ML engineers. The short-term target is robots that support skilled workers building infrastructure; the long-term goal is personal robots [5].
- **Nemotron 3 Ultra gave NVIDIA a larger US open-weight contender.** According to Artificial Analysis, the 550B-parameter / 55B-active, 90%-sparse model is the largest Nemotron 3 release and the most

intelligent US open-weights model so far, scoring 48 on its Intelligence Index and serving above 300 tokens/sec on a pre-release DeepInfra endpoint [6]. Additional benchmarks are still to come at release [6].

Research & Innovation

Why it matters: the most useful technical work focused on physical AI, training efficiency, and cheaper long-context agents.

- **Cosmos 3 unifies reasoning, world modeling, and action generation for physical AI.** NVIDIA's new architecture replaces separate perception, prediction, and action models with a two-part system—Reasoner Tower and Generator Tower—aimed at robots, autonomous vehicles, and smart environments [7]. The architecture uses Mixture-of-Transformers [8].
- **DiffusionBlocks targets training memory, not just model quality.** Sakana AI says it can train networks one block at a time, drastically reducing memory requirements while remaining competitive with end-to-end training across ViT, DiT, masked diffusion, autoregressive transformers, and recurrent-depth transformers; code is already out for ViT [9, 10].
- **The Efficiency Frontier paper reframes context management as an optimization problem.** It models retrieval, compression, and full-context prompting under a single cost-performance objective; on 5,000 HotpotQA examples, deployment-aware selection cut effective token use by about 25%, and amortized memory compression was over 50% cheaper than full-context prompting in higher-performance settings [11].

Products & Launches

Why it matters: launches were strongest in multimodal creation, long-video understanding, and agent usability.

- **HiDream O1 Image arrived as an open-source image stack with strong arena results.** The family spans three open-weight models for text-to-image and instruction-based editing, with the base and Dev models accepting text plus up to 10 images. Artificial Analysis said Dev-2604 leads open-weight models on its Text-to-Image Arena, and the weights plus full inference pipeline are released under MIT [12].
- **Keye-VL-2.0 brought sparse attention into long-video multimodality.** ModelScope said the 30B-A3B release is the first multimodal model with DeepSeek Sparse Attention, supports a 256k context window for hour-long video processing, and outperforms 200B+ open models on LongVideoBench while cutting prefill costs by 50% [13].
- **Hermes Agent now has native Windows support.** Nous Research said Windows support is out of beta, installable directly from PowerShell,

extending the full desktop agent experience to Windows users [14, 15, 16].

Industry Moves

Why it matters: today's bigger strategic moves were about compute, platform control, and ecosystem alignment.

- **NVIDIA said Vera Rubin is entering full production for agentic AI factories.** The company described it as a POD-scale platform for agentic workloads with end-to-end security, backed by Taiwan server makers and a broader manufacturing, cloud, and infrastructure ecosystem [17].
- **Apple's AI stack may be shifting toward Google and NVIDIA infrastructure.** A post citing *The Information* said Apple's upcoming Siri and on-device AI upgrade centers on a distilled Gemini model running locally on iPhone silicon, while heavier queries route to Google Cloud using NVIDIA confidential-compute technology—a change from Apple's earlier Private Cloud Compute positioning [18].
- **Nous Research is aligning Hermes Agent with NVIDIA's new edge stack.** Nous said it has been working with NVIDIA so Hermes Agent runs on RTX Spark and integrates with OpenShell, which connects Hermes to Microsoft security primitives [19].

Quick Takes

Why it matters: a few smaller updates sharpened the picture on evals, local inference, and hands-on agent behavior.

- **Claude 4.8 Opus** set a new high on GBA Eval, where models build a working Game Boy Advance emulator within 24 hours [20].
- **OBLIQ-Bench** was proposed as a harder IR benchmark after frontier-LLM oracle reranking showed little headroom on older search benchmarks [21, 22, 23].
- **vLLM + DGX Spark** showed desk-side large-model inference with streaming responses, paged KV cache, runtime tuning, and Prometheus metrics [24].
- **A Codex user reported** that, given only two MP4 filenames, Codex found the files, verified codec, fps, and resolution, stitched them without re-encoding, and spot-checked the transition [25].

Sources

1. X post by @MiniMax_AI
2. X post by @OpenRouter
3. X post by @ollama

4. X post by @togethercompute
5. X post by @sama
6. X post by @ArtificialAnlys
7. X post by @TheTuringPost
8. X post by @TheTuringPost
9. X post by @SakanaAILabs
10. X post by @SakanaAILabs
11. X post by @omarsar0
12. X post by @ArtificialAnlys
13. X post by @ModelScope2022
14. X post by @Teknium
15. X post by @NousResearch
16. X post by @Teknium
17. X post by @nvidianewsroom
18. X post by @kimmonismus
19. X post by @NousResearch
20. X post by @scaling01
21. X post by @dianetc_
22. X post by @lateinteraction
23. X post by @lateinteraction
24. X post by @vllm_project
25. X post by @reach_vb