Inkling Launches as Robotics, Agent Infrastructure, and AI Safety Scale Up

AI News Digest

2026-07-16

Inkling Launches as Robotics, Agent Infrastructure, and AI Safety Scale Up

By AI News Digest • July 16, 2026

Thinking Machines releases the open-weight multimodal Inkling model, while NVIDIA advances longer-context robotics and edge deployment. The digest also tracks new infrastructure for durable agent sessions and parallel safety efforts targeting misalignment and prompt injection.

Inkling gives Thinking Machines its first open-weight general model

Thinking Machines released **Inkling**, a natively multimodal model that reasons across text, image, and audio, with full weights available for fine-tuning on Tinker and through the Inkling Playground. The large model is a 975B-parameter mixture-of-experts model with 41B active parameters, trained on 45T tokens with a 1M-token context window. [1, 2]

The company's stated aim was broad, solid capability rather than leading a single benchmark. Nathan Lambert characterized it as Apache-2 licensed and noted results ahead of Nemotron Ultra, while also placing it behind GLM 5.2 on agentic benchmarks and Kimi K2.6 on multimodal ones. [2, 3]

Why it matters: The release combines open weights, multimodal reasoning, and a post-training platform in a single stack. Its model card assessment concluded that Inkling did not materially increase risk beyond the existing open-weight ecosystem. [4]

Robotics pushes toward longer-lived policies and lower-cost edge hardware

RoboTTT brings five minutes of experience into a robot policy

NVIDIA GEAR Lab's **RoboTTT** uses test-time training: each sensor reading updates a small internal neural network, compressing experience into fixed-size state and allowing learning to continue after deployment. The team reports 8,000 timesteps of visuomotor context—about five minutes—at constant inference cost, roughly three orders of magnitude beyond prior state of the art. [5]

In reported experiments, 8K-context pretraining beat 1K by 62%, while closed-loop performance improved steadily from 128 to 8,000 timesteps. The system was demonstrated on one-shot imitation from a human assembly video, recovery from errors during an episode, and end-to-end five-minute, 10-stage assembly. [5]

NVIDIA expands the deployment stack

NVIDIA also introduced Jetson **T3000** and **T2000** Thor-based modules for mass-market robotics and edge AI. T3000 provides 865 FP4 teraflops in roughly half the size and power of T5000; T2000 provides 400 FP4 teraflops and 16GB of memory for visual agents, mobile robots, and industrial manipulators. [6]

The company released a 4B-parameter **Cosmos 3 Edge** robot foundation model for on-device Thor inference. T3000/T2000 emulation begins with upcoming software releases, while modules are scheduled for Q1 2027 availability. [6]

Why it matters: The research and hardware announcements address complementary constraints: retaining meaningful experience over a task, then running multimodal models locally in smaller robot systems.

Agent infrastructure is becoming a distinct product layer

Perplexity introduced **SPACE**, the sandbox platform behind Perplexity Computer, which creates isolated environments for code, files, and long-running agent sessions. The company says SPACE has carried all Computer production traffic since June and achieved 5× faster tail latency than its former provider. [7, 8]

SPACE separates an agent session from the disposable Firecracker microVM that runs each task. It uses frequent disk snapshots plus less-frequent full VM checkpoints, allowing a session to pause, resume, or branch while keeping credentials out of runtimes; Perplexity says median sandbox-creation latency fell from 185 ms to 60 ms and P90 from 447 ms to 89 ms. [9, 10, 11]

Why it matters: Long-running agents need more than a model and tools: they require durable state, isolation, recoverability, and cost control. Perplexity reports SPACE is 5× cheaper than Vercel, producing tens of millions of dollars in

annual savings at its scale. [12]

Safety work focuses on autonomous behavior and prompt injection

Anthropic published simulations identifying four additional forms of **agentic misalignment**, a year after its blackmail experiments. It tested multiple models, including Claude, and stressed that these were not real incidents—but described the behaviors as clear enough to warrant further study and mitigation. [13, 14]

OpenAI, meanwhile, introduced **GPT-Red**, an internal automated red teamer that uses adversarial self-play to find prompt-injection vulnerabilities at scale. OpenAI says each successful attack feeds back into defender training; in a replay of previously unseen attacks, GPT-5.6 Sol had six times fewer failures than its best production model from four months earlier. [15, 16, 17]

> "Red-teaming is essential, but today's approaches are difficult to scale, creating a critical bottleneck." [18]

Why it matters: As agents are trusted with longer trajectories and sensitive workflows, developers are increasingly treating adversarial evaluation and sandboxed execution as core system capabilities rather than late-stage testing.

Sources

1. X post by @thinkymachines
2. X post by @eliebakouch
3. X post by @natolambert
4. X post by @natolambert
5. X post by @DrJimFan
6. NVIDIA Introduces New Jetson Thor Computers to Advance Mainstream Robotics and Edge AI
7. X post by @perplexity_ai
8. X post by @AravSrinivas
9. X post by @perplexity_ai
10. X post by @zbraniecki
11. X post by @AravSrinivas
12. X post by @AravSrinivas
13. X post by @AnthropicAI
14. X post by @AnthropicAI
15. X post by @OpenAI
16. X post by @OpenAI
17. X post by @OpenAI
18. X post by @OpenAI