OpenAI Expands Real-Time Agents as Training-Time Speedups Stack Up

AI News Digest

2026-05-14

OpenAI Expands Real-Time Agents as Training-Time Speedups Stack Up

By AI News Digest • May 14, 2026

OpenAI pushed voice and coding agents closer to production, while Nous, Lighthouse Attention, and Adaption highlighted a notable wave of training-time efficiency ideas. The day also brought major infrastructure signals around large-scale reinforcement learning and Europe's compute buildout.

The clearest pattern

Today's strongest thread is practical AI: companies are making agents easier to ship, while researchers are looking for training-time optimizations that preserve a conventional inference-time model [1, 2, 3, 4].

Agents move closer to production

OpenAI broadens its real-time voice stack

OpenAI said last week's audio release included a real-time translation model with 70+ input and 13 output languages, a GPT real-time Whisper model with 80 input languages and latency as low as 200 ms, and GPT Realtime 2, which it described as its most intelligent voice model [2]. GPT Realtime 2 adds a 128k-token context window, parallel tool calls, dynamic voice cloning and tone matching, controllable expressiveness, and stronger domain vocabulary and tool calling; demos showed it operating an e-commerce UI and a product analytics dashboard through tools [2]. In Sierra's early testing, calls were roughly 30% faster at P50 and up to 200% faster at P90 than its cascaded system, though Sierra emphasized that production use still depends on a separate harness for workflows, guardrails, redaction, and policy control [2].

Why it matters: The release pushes voice systems further from transcription-and-response toward voice agents that can act inside software and production workflows [2].

Build Hour: GPT-Realtime-2 (6:57)

Codex demand is arriving alongside sandboxing

OpenAI said 2,000 developers reached out about Codex in three hours, and Greg Brockman reported strong enterprise excitement about adopting it [5, 6]. Sam Altman also offered companies two months of free Codex usage for the next 30 days if they want to try switching over [7]. Separately, OpenAI published how it built a Windows sandbox for Codex to avoid forcing developers to choose between constant approval prompts and full machine access [1].

Why it matters: The product signal here is paired with a deployment signal: security boundaries are becoming part of the coding-agent product itself, not an afterthought [1, 6].

Training-time gains without inference-time changes

Nous reports 2–3x wall-clock gains from Token Superposition Training

Nous Research released Token Superposition Training, a pretraining modification that it says delivers a 2–3x wall-clock speedup at matched FLOPs without

changing model architecture, optimizer, tokenizer, or training data [3]. For the first third of training, the model reads and predicts contiguous bags of tokens with averaged input embeddings and a modified output loss, then returns to standard next-token prediction for the rest of the run [3]. Nous said it validated the approach at 270M, 600M, and 3B dense scales, plus a 10B-A1B MoE [3].

Why it matters: The notable claim is not only the speedup, but that the final inference-time model remains identical to one produced by conventional pre-training [3].

Lighthouse Attention aims to speed long-context training, then disappear

Lighthouse Attention wraps standard SDPA with a hierarchical, gradient-free selection layer that compresses and decompresses queries, keys, and values while preserving left-to-right causality [4]. The method can be removed near the end of training through a short recovery phase, and preliminary LLM experiments reported faster total training time and lower final loss than full-attention baselines [4, 8]. Sebastian Raschka highlighted this as a relatively low-commitment attention modification because teams can switch back to vanilla attention near the end and recover roughly the same modeling performance [8].

Why it matters: Like TST, this is a training-time efficiency idea aimed at avoiding deployment-time architectural cost [4, 8].

AutoScientist turns the training research loop into a product

Adaption launched AutoScientist to automate the full research loop for model training, arguing that most model training fails outside frontier labs and that even inside them, many training choices are still a matter of taste [9, 10]. Sara Hooker said AutoScientist beat hand-engineered configs from Adaption's research staff across verticals, model types, and dataset sizes, with consistent results and more predictable performance [11, 12]. She also framed the result as important for AI progress outside a small number of proprietary labs [13].

Why it matters: If the research loop around training becomes automatable, some model-development advantage could shift from tacit tuning intuition toward searchable, repeatable systems [10, 13].

The infrastructure conversation is widening

NVIDIA and David Silver's Ineffable Intelligence are building for large-scale RL

NVIDIA and Ineffable Intelligence, the London lab founded by David Silver, said they are collaborating to build infrastructure for large-scale reinforcement learning [14]. NVIDIA said RL workloads generate data on the fly in tight act-observe-score-update loops that stress interconnect, memory bandwidth, and

serving differently from pretraining, and that the work starts on Grace Blackwell while exploring Vera Rubin [14].

> "The next frontier of AI is superlearners — systems that learn continuously from experience." [14]

Why it matters: This is a concrete industry bet that scaling RL will require its own hardware-software pipeline, not just a larger version of pretraining infrastructure [14].

Mistral makes Europe's compute challenge unusually explicit

In testimony to the French National Assembly, Arthur Mensch said Mistral now has 1,000 employees, a €12B valuation, and a €1B revenue target by year-end [15]. He warned that the decisive window is the next two years because supply could be locked up before Europe builds enough capacity, and projected AI demand on the order of 1 kW per person within five years — implying roughly 40 GW for France and 400 GW for Europe [15].

Why it matters: The European AI debate is being framed less as a pure model contest and more as an energy, capital, and industrial-capacity problem [15].

Sources

1. X post by @OpenAIDevs
2. Build Hour: GPT-Realtime-2
3. X post by @NousResearch
4. X post by @omarsar0
5. X post by @OpenAIDevs
6. X post by @gdb
7. X post by @sama
8. X post by @rasbt
9. X post by @adaption_ai
10. X post by @sarahookr
11. X post by @sarahookr
12. X post by @sarahookr
13. X post by @sarahookr
14. NVIDIA, Ineffable Intelligence Team Up to Build the Future of Reinforcement Learning Infrastructure
15. Arthur Mensch, cofondateur de Mistral AI, est auditionné à l'Assemblée nationale - 12/05/2026