Codex Goes Mobile, Figure Extends Humanoid Runtime, and Autonomous Agents Beat a Human Baseline

AI High Signal Digest

2026-05-15

Codex Goes Mobile, Figure Extends Humanoid Runtime, and Autonomous Agents Beat a Human Baseline

By AI High Signal Digest • May 15, 2026

Codex went mobile, Figure extended humanoid runtime past a full day, and PrimeIntellect showed autonomous coding agents beating a human nanoGPT baseline. The brief also covers diffusion decoding speedups, time-series scaling laws, enterprise data agents, Anthropic's Gates partnership, and the latest U.S.-China compute tensions.

Top Stories

Why it matters: today's strongest signal is that AI agents are becoming more persistent, more physical, and more capable at open-ended technical work.

- **OpenAI put Codex on the phone.** Codex is now in preview inside the ChatGPT mobile app, letting users start work, review outputs, steer execution, and approve next steps from iOS and Android while jobs keep running on a laptop, Mac mini, or devbox; OpenAI also made Remote SSH generally available for managed remote environments [1, 2, 3]. Commentators called it a major unlock for remote agent work and broader day-to-day agent usage [4, 5].
- **Figure pushed humanoid uptime from a shift demo to around-the-clock operation.** Figure said its F.03 robots moved from an original 8-hour target to more than 24 hours of continuous autonomous package sorting without failure, and later crossed 30 hours with no downtime [6, 7]. The company says the robots are now around human parity at roughly

3 seconds per package, run entirely onboard via Helix-02 with no teleoperation, and have processed more than 38,000 packages [6, 7].

- **Autonomous coding agents beat the human baseline on nanoGPT optimization.** PrimeIntellect let Claude Code (Opus 4.7) and Codex (GPT-5.5) run autonomously on the nanoGPT speedrun optimizer track using idle compute, totaling about 10,000 runs, 14,000 H200 hours, and 23.9B tokens [8, 9]. Opus reached 2930 steps and Codex 2950, both ahead of the 2990 human baseline; PrimeIntellect framed the work as a step toward automating AI research [8, 9].

Research & Innovation

Why it matters: the most notable technical updates were about cheaper inference, clearer scaling laws, and better understanding of what models are doing internally.

- **Zyphra's diffusion language model targets the decoding bottleneck.** ZAYA1-8B-Diffusion-Preview, trained on AMD hardware, claims a 4.6-7.7x decoding speedup over autoregressive LLMs with minimal quality degradation by generating 16-token blocks in parallel [10, 11, 12]. The company argues this matters because autoregressive inference is memory-bandwidth bound, while diffusion removes that bottleneck [13].
- **Datadog's Toto 2.0 makes the case that time-series models scale cleanly.** The open-weights family ranges from 4M to 2.5B parameters, with each size outperforming the previous one under a single hyperparameter configuration and leading BOOM, GIFT-Eval, and TIME [14]. Datadog's framing is that time series now shows the kind of predictable scaling behavior long seen in language and vision [14].
- **Goodfire found a "geometric calculator" inside Llama models.** The mechanism encodes numbers as positions on multiple circles, handles arithmetic as well as weekday and month reasoning, and was tested by steering the circles and watching answers change [15, 16, 17]. Goodfire says this kind of neural-geometry work could improve debugging, control, and model design [18].

Products & Launches

Why it matters: new tools keep turning agents from isolated assistants into systems that can work across design, data, and the browser itself.

- **MagicPath 2.0** is now a multiplayer canvas for humans and agents such as Codex and Claude Code, with real-time shared context and fully functional browser-based prototypes built from real code [19, 20, 21]. It also supports design-to-repo and repo-to-design round trips through external agents [22].

- **Perplexity Computer now connects to Snowflake.** The product can run end-to-end workflows on live warehouse data and return answers with SQL, source tables, filters, and metrics, while admins retain control over access and shared business logic [23, 24].
- **Kimi Web Bridge brings browser actions to major agent stacks.** The extension lets agents search, scroll, click, type, fill spreadsheets, and turn repeated browser work into reusable skills; it supports Kimi Code CLI, Claude Code, Cursor, Codex, Hermes, and more [25, 26, 27].

Industry Moves

Why it matters: major firms are pairing frontier models with real distribution, public-interest deployment, and international expansion.

- **Anthropic partnered with the Gates Foundation on a $200M package** of grants, Claude credits, and technical support across global health, life sciences, education, agriculture, and economic mobility [28].
- **Runway is expanding to Japan with a Tokyo base.** The company says Japan is already its third-largest market, its fastest-growing self-serve market in Asia, and has seen 300% enterprise customer growth over the last 12 months [29, 30].

Policy & Regulation

Why it matters: AI geopolitics still turns on compute, and approvals matter less than actual hardware movement.

- **U.S.-China chip controls remain unresolved in practice.** Reuters-reported approvals cover roughly 10 Chinese firms buying Nvidia H200s, but no chips have shipped yet [31]. Separate analysis this week argued Chinese labs remain compute-constrained and continue renting or smuggling Nvidia-designed chips from third countries, so the real signal is deliveries, not approvals [32, 33].

Quick Takes

Why it matters: these smaller updates point to where the next wave of tooling, governance, and specialty models is heading.

- Ahead of Google I/O, a leak described **Gemini Spark** as an always-on agent with access to Gmail, Calendar, location, tasks, and personal context [34].
- **arXiv** now has a **one-year ban** for hallucinated references in submissions [35].
- **Baseten** says it serves **Qwen3-TTS** on **vLLM-Omni** at **$3 per 1M characters**, about **90% lower** than comparable closed-source TTS APIs [36, 37].

- **Intern-S2-Preview**, a **35B** open scientific multimodal model, claims performance comparable to the trillion-scale Intern-S1-Pro on core scientific tasks and launched with day-0 vLLM support [38, 39].

Sources

1. X post by @OpenAI
2. X post by @OpenAI
3. X post by @OpenAIDevs
4. X post by @gdb
5. X post by @thursdai_pod
6. X post by @adcock_brett
7. X post by @adcock_brett
8. X post by @PrimeIntellect
9. X post by @eliebakouch
10. X post by @ZyphraAI
11. X post by @ZyphraAI
12. X post by @ZyphraAI
13. X post by @ZyphraAI
14. X post by @ClementDelangue
15. X post by @GoodfireAI
16. X post by @GoodfireAI
17. X post by @GoodfireAI
18. X post by @GoodfireAI
19. X post by @skirano
20. X post by @skirano
21. X post by @skirano
22. X post by @skirano
23. X post by @perplexity_ai
24. X post by @perplexity_ai
25. X post by @Kimi_Moonshot
26. X post by @Kimi_Moonshot
27. X post by @Kimi_Moonshot
28. X post by @AnthropicAI
29. X post by @nikkei
30. X post by @c_valenzuelab
31. X post by @dnystedt
32. X post by @RyanFedasiuk
33. X post by @kimmonismus
34. X post by @kimmonismus
35. X post by @stuhlmueller
36. X post by @baseten
37. X post by @vllm_project
38. X post by @intern_lm

39. X post by @vllm_project