Codex Becomes a Broader Work Surface as AI Efficiency and Agent Evaluation Advance

AI News Digest

2026-07-15

Codex Becomes a Broader Work Surface as AI Efficiency and Agent Evaluation Advance

By AI News Digest • July 15, 2026

OpenAI turns Codex into a more complete agentic work environment, while Perplexity releases a benchmark for wide research. The digest also tracks compression and inference-efficiency claims, the growing emphasis on agent oversight and evaluation, and ChaiDiscovery's major funding round.

OpenAI folds Codex into ChatGPT and expands its agentic work surface

GPT-5.6 Sol powers a broader coding-and-work environment

OpenAI has brought Codex into ChatGPT in a dedicated space alongside a new work agent, while making its GPT-5.6 Sol frontier model available to all users. [1] Codex can now automatically divide work across subagents, operate browsers—including apps requiring logins and passkeys—and deploy full-stack applications through Sites with hosting, authentication, database, and storage built in. [1]

OpenAI also reported that usage of its agentic products—Codex and ChatGPT Work—rose 2.5× over the past week. [2]

Why it matters: The release packages model reasoning, delegated coding work, browser interaction, deployment, and collaboration into a single workflow rather than treating code generation as a standalone feature.

Perplexity opens a benchmark for wide research agents

WANDR targets real knowledge-work tasks that remain difficult for frontier models

Perplexity has open-sourced WANDR, an internal benchmark it used to develop deep and wide research capabilities. The suite contains 500 tasks built on real knowledge work and is described as difficult even for today's most powerful models. [3, 4]

Alongside the benchmark, Perplexity made a Wide Research preset available in its Agent API. It says its "Search as Code" architecture lets a model design research once and execute it deterministically at scale without overwhelming context. [5, 4]

Why it matters: Research agents need evaluation beyond isolated question answering; WANDR offers a shared target for measuring systems that must explore broadly as well as reason deeply.

Efficiency advances span model compression and datacenter inference

Tencent reports a 1-bit, 295B model that fits in 88GB

Tencent released 1-bit and 4-bit quantized versions of its 295B Hy3 model, saying it can be served on a single GPU with llama.cpp and MTP enabled. [6] Emad Mostaque reported 75.4% on SWE-Bench Verified and 53.9% on SWE-Bench Pro for the 1-bit version, with an 88GB footprint; he characterized the drop from 16-bit precision as roughly 5%. [7, 8]

At the infrastructure layer, NVIDIA says its GB300 NVL72 systems deliver up to 25× Hopper's performance per watt on DeepSeek V4 Pro, 20× on GLM5.1, and 10× on Kimi K2.6. [9] NVIDIA also says Anthropic and OpenAI use Blackwell NVL72 systems for inference, while CoreWeave, Perplexity, and Fireworks AI run open models in production on the platform. [9]

Why it matters: Compression and rack-scale efficiency are advancing in parallel, potentially changing where capable models can run and the economics of serving them.

The engineering focus shifts from agents to the systems around them

Reliability, evaluation, and human oversight take center stage

A Latent.Space review of AI Engineer World's Fair 2026 reports a shift from prompting individual agents toward building "harnesses" that manage workflows, context, permissions, evaluation, persistent state, and continuous improvement. [10] Speakers framed this as "loop engineering": agents can run

an inner execution loop, while people retain an outer loop for feedback, evaluations, direction, and decisions. [10]

François Chollet similarly argued that enterprises need stronger evaluations for knowledge-work workflows, so they can tell whether a model, prompt, or system change improved or broke performance. [11]

Why it matters: As agentic tools move into longer-running work, the differentiator is increasingly the operational system that constrains, measures, and improves them—not only the underlying model.

ChaiDiscovery raises $400 million for AI drug discovery

New capital follows deployments with major pharmaceutical companies

ChaiDiscovery has raised a $400 million Series C at a $3.8 billion valuation from Index Ventures, Kleiner Perkins, Sequoia, and Dimension Capital. [12] The company has deployments with Eli Lilly, Pfizer, and Novartis, according to Sarah Guo. [13]

Why it matters: The financing is a substantial vote of confidence in AI-enabled drug discovery, paired with reported use at several large pharmaceutical companies.

Sources

1. Codex just got better for developers
2. X post by @sama
3. X post by @perplexity_ai
4. X post by @perplexitydevs
5. X post by @AravSrinivas
6. X post by @TencentHunyuan
7. X post by @EMostaque
8. X post by @EMostaque
9. Why Performance per Watt Is the Ultimate Metric for AI Infrastructure Efficiency
10. 5 Trends That Defined AI Engineering at World's Fair 2026
11. X post by @levie
12. X post by @joshim5
13. X post by @saranormous