Qwen 3.7-Max Raises the Agent Bar as Codex Expands Computer Use

AI High Signal Digest

2026-05-22

Qwen 3.7-Max Raises the Agent Bar as Codex Expands Computer Use

By AI High Signal Digest • May 22, 2026

Alibaba's Qwen3.7-Max narrowed the gap to frontier labs with stronger agentic performance, while OpenAI pushed Codex further into persistent computer use. The brief also covers new research on AI review quality, long-context architectures, product launches from Cohere and Devin, and the latest business signals across AI infrastructure.

Top Stories

Why it matters: the clearest signals today were stronger agentic models, faster productization of computer-use systems, and continued compression in AI price-performance.

- **Alibaba pushed deeper into the frontier with Qwen3.7-Max.** The company introduced it as a flagship model for the "Agent Era," highlighting end-to-end coding, MCP-based productivity workflows, and a 35-hour kernel-optimization run that used 1,158 tool calls and achieved a 10.0x geometric-mean speedup over the Triton reference [1, 2]. Artificial Analysis scored it at **56.6**, up **4.8 points** from Qwen3.6 Max Preview and the closest Alibaba has come to frontier labs, with gains concentrated in scientific reasoning, agentic capability, and coding [3]. Part of that gain came from higher abstention that reduced hallucination rate, not just higher factual recall [3].
- **Model quality keeps getting cheaper.** Text Arena's price-performance view says the cost of frontier-quality output fell from about **$50** per million tokens in 2023 to about **$0.10** today, while the gap between sub-$0.20 models and the leader shrank from roughly **350** Arena points to **60** [4]. In coding agents, Cursor's **Composer 2.5** reached **62**

on the Artificial Analysis Coding Agent Index—third overall—at **$0.07** per task in standard mode or **$0.44** in Fast mode, versus **$4.10-$4.82** for the two higher-ranked systems [5].

Research & Innovation

Why it matters: the most useful research updates were about better evaluation, better long-context architectures, and better harnesses rather than just bigger models.

- **AI paper review got a strong benchmark result.** A study on **82 Nature-family papers** found that frontier LMs in an agent harness were judged by **45 expert scientists** to outperform the best human reviewer; the authors also said AI reviews were accurate and well-evidenced but less grounded in scientific norms and more homogeneous than human panels [6, 7].
- **Gated DeltaNet-2 advanced linear attention.** The architecture decouples erase and write gates, outperformed KDA and Mamba-3 at **1.3B** scale, and showed especially large long-context gains, including **S-NIAH-3: 63 → 90** and **multi-key needle retrieval: 28 → 38** [8].
- **Harness quality still matters.** The new **Physics-Intern** scaffold wraps a model with a dedicated subagent for science problems; it raised Gemini 3.1 Pro from **17.7** to **31.4**, beating GPT 5.5 Pro, while GPT 5.5 Pro itself did not improve under the harness [9].

Products & Launches

Why it matters: leading products are moving from chat and code generation toward persistent work, computer use, and lower-cost deployment.

- **OpenAI expanded Codex into a more persistent computer-use product.** New updates let Codex securely use apps on a locked Mac from a phone, run in **Goal mode** across the app, IDE extension, and CLI for tasks lasting hours or days, pull screenshots plus visible and off-screen text into threads via **Appshots**, and make direct webpage changes with advanced annotation [10, 11, 12, 13]. OpenAI also added richer business analytics and team plugin sharing [14, 15].
- **Cohere open-sourced Command A+.** Cohere called it its most powerful LLM yet, optimized to run on minimal hardware and released for broad access; separate posts described it as the company's first fully open-source **Apache 2** model, and a **W4A4** Hugging Face release promises sharply lower serving footprint with little performance loss [16, 17, 18].
- **Devin gained native Windows support.** Cognition said Devin can now run in a Windows VM with support for MSBuild, IIS, PowerShell,

SQL Server, and enterprise controls including isolated sessions, SOC 2 Type II, ISO 27001, SSO, and RBAC [19, 20].

Industry Moves

Why it matters: the business story is increasingly about who can convert model demand into durable revenue, infra scale, and profitable software.

- **The frontier revenue race is separating into growth and profitability stories.** Posts citing reported figures put OpenAI at about **$5.7B** in Q1 revenue versus Anthropic at roughly **$4.7B-$4.8B** [21, 22, 23]. But Anthropic's recent annualized revenue reportedly neared **$45B** and it is projecting about **$600M** in operating profit, while separate commentary said OpenAI was losing **$1.22** for every dollar earned and had user growth stalled near **905M** weekly actives [21, 22, 23].
- **Modal raised $355M at a $4.65B valuation.** The round was led by General Catalyst and Redpoint, with the company framing its mission around infrastructure that improves developer productivity for AI and data teams as workloads scale [24, 25].
- **turbopuffer reported breakout traction in search infrastructure.** The company said it crossed **$100M run-rate** in March, just 19 months after $1M ARR, while staying profitable with under $1M raised; it says customers include Cursor, Anthropic, and Cognition, and that Cursor cut costs **95%** after migrating production search workloads [26, 27].

Quick Takes

Why it matters: a few smaller updates sharpened the picture on chips, local AI, robotics, and startup distribution.

- **HBM is eating more of AI chip budgets:** its share of frontier chip component spend rose from **52% to 63%**, with total spend growing from about **$12B to $32B** [28, 29].
- **llama.cpp added WebGPU support,** enabling GPU-accelerated local models in the browser with no data leaving the device [30].
- **Figure said F.03 reached 200 hours of autonomous operation without failure** after processing **238,000 packages** [31, 32].
- **OpenAI is offering $2M in tokens to every YC company** in the spring and summer batches [33].

Sources

1. X post by @Alibaba_Qwen
2. X post by @Alibaba_Qwen
3. X post by @ArtificialAnlys

4. X post by @arena
5. X post by @ArtificialAnlys
6. X post by @seungonekim
7. X post by @gneubig
8. X post by @ahatamiz1
9. X post by @lvwerra
10. X post by @OpenAI
11. X post by @OpenAI
12. X post by @OpenAIDevs
13. X post by @OpenAI
14. X post by @OpenAIDevs
15. X post by @OpenAIDevs
16. X post by @cohere
17. X post by @aidangomez
18. X post by @cohere
19. X post by @cognition
20. X post by @cognition
21. X post by @wallstengine
22. X post by @kimmonismus
23. X post by @srimuppidi
24. X post by @bernhardsson
25. X post by @sarahcat21
26. X post by @Sirupsen
27. X post by @Sirupsen
28. X post by @EpochAIResearch
29. X post by @EpochAIResearch
30. X post by @reeselevine
31. X post by @Figure_robot
32. X post by @adcock_brett
33. X post by @ycombinator