Cursor Cloud Agents go video-first + test-first, while GPT-5.4 upgrades Codex and always-on automations spread

Coding Agents Alpha Tracker

2026-03-06

Cursor Cloud Agents go video-first + test-first, while GPT-5.4 upgrades Codex and always-on automations spread

By Coding Agents Alpha Tracker • March 6, 2026

Cursor's Cloud Agents show what "agentic IDE design" looks like in practice: dedicated VMs, end-to-end testing, demo videos, and Slack-first collaboration. Plus: GPT-5.4's Codex upgrades (/fast mode, Playwright skill, 1M context status), always-on Cursor Automations, and hard lessons on evaluation, manual testing, and CI prompt-injection security.

TOP SIGNAL

Cursor's latest **Cloud Agents** push is a concrete "agentic IDE" redesign: agents run in **dedicated VMs**, **test changes end-to-end**, and return a **demo video + a tested PR**, with **remote desktop/terminal** access for quick human iteration [1][2]. Cursor says this flow exists because **reviewing code becomes the bottleneck** once agents can generate large diffs—video is an easier first review surface (but not a code-review replacement) [3].

TOOLS & MODELS

- **OpenAI — GPT-5.4 rollout (Thinking + Pro), unified frontier model**

[1] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor

[2] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor

[3] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor

  - Rolling out in **ChatGPT**, and also available in the **API and Codex** [4].
  - OpenAI describes it as bringing advances in **reasoning, coding, and agentic workflows** into one model [5].
  - Practitioner note: Hanson Wang says **Codex and Thinking models are now unified** [6].
- **Codex — `/fast` mode (GPT-5.4)**
  - Claimed **1.5x faster** with "the same intelligence and reasoning" [7].
  - Tradeoff called out by the Codex team: **1.5x speed for 2x cost** [8].
- **Codex — Playwright skill + frontend improvements (GPT-5.4 era)**
  - Romain Huet says complex frontend work looks "noticeably better," and calls out a new **Playwright skill** that lets Codex **visually debug and test apps while it builds** [9].
- **Cursor — GPT-5.4 support + 1M context status**
  - Cursor says GPT-5.4 is now available and is "more natural and assertive," leading on their internal benchmarks [10].
  - Cursor's Jediah Katz reported an issue with **1M context** in GPT-5.4 and said they were fixing it ASAP [11].
  - Follow-up: Katz says **1M context is now available** for GPT-5.4 if you toggle **Max Mode** on [12] (enterprise legacy pricing: coming behind a separate **gpt-5.4-1m** slug [13]).
- **Cursor — Automations (always-on agents)**
  - Cursor announced **Automations**: "continuously monitor and improve your codebase," running on **triggers and instructions you define** [14].
  - Cursor CEO Michael Truell says Automations already run **thousands of times per day** internally, powering **self-healing CI**, **auto-approving PR flows**, **compute-intensive security review**, and a **team-wide memory system** [15].
  - Jediah Katz highlights they can trigger on **any event/webhook**, run **in the cloud** (not dependent on one laptop), and are **team-owned** [16].
- **Local agents (privacy-driven) — Qwen 3.5 as "good enough" for

[4] post by @OpenAI

[5] post by @OpenAI

[6] post by @hansonwng

[7] post by @OpenAIDevs

[8] post by @embirico

[9] post by @romainhuet

[10] post by @cursor_ai

[11] post by @jediahkatz

[12] post by @jediahkatz

[13] post by @jediahkatz

[14] post by @cursor_ai

[15] post by @mntruell

[16] post by @jediahkatz

some tasks
  - Salvatore Sanfilippo says Qwen 3.5 is the first time he feels local agents can work for **simpler programming tasks** on your own machine (not state of the art, but effective) [17].
  - He compares the **27B dense** model (more stable, good for GPU) and **35B MoE (3B active)** (faster iteration, maybe better in practice) [18].
- **Augment — "Intent" UI for large workloads**
  - Theo describes Intent as a shift from chat/autocomplete toward a UI for **planning and managing large agentic coding workloads** [19].
  - He also highlights pulling context from **Linear, Sentry, GitHub issues, or PRs** to keep workstreams compatible [20].

WORKFLOWS & TRICKS

1) Cursor's "Cloud Agent" loop (test-first + video-first + HITL)

A replicable loop Cursor describes for cloud-agent work: - **Kick off an agent in cursor.com/agents**; it works longer because it **tests end-to-end** (starts dev servers, iterates) and aims to return a **tested PR** [21]. - First review pass: **watch the demo video** (a faster entry point than reviewing a huge diff) [22]. - If needed: use **remote desktop (VNC-style) + terminal access** to interactively verify behavior and iterate [23]. - Testing controls: - Default behavior is calibrated testing: don't test "very simple copy changes," but test complex ones; configurable via **agents.md** [24]. - Use `/notest` to force skipping tests [25].

2) Bugfixes that ship faster: `/repro` before/after videos

Cursor's `**/repro**` pattern: - Agent **reproduces the bug** and records a video, then **fixes** and records an "after" video [26]. - Cursor says this moves some bug classes from "hard to repro locally" to "merge in ~90 seconds" [27].

3) Parallelism you can actually review: Best-of-N via 20s videos

- Cursor says demo videos made them use **best-of-N** more often because reviewing **four 20-second videos** is manageable vs reviewing **4× giant**

[17] Qwen 3.5
[18] Qwen 3.5
[19] The drama never ends…
[20] The drama never ends…
[21] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[22] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[23] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[24] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[25] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[26] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[27] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor

diffs [28].

4) Slack as the "new IDE" surface (team workflows)

- Cursor engineers describe Slack threads as a dev surface: you can **@cursor** in issue/product channels to kick off a cloud agent; teammates can "follow up" in-thread with more context [29].
- They say the human discussion shifts to the high-order decisions ("do we ship this?", "is this the right UX?") while the agent handles implementation [30].

5) Subagents for context + compute management

- Cursor highlights **subagents** as a way to delegate across prompts/goals/models and keep context manageable [31].
- Example: an **explore** subagent can be routed to a faster model to read lots of code quickly, then summarize back to the parent agent [32].

6) Long-running agent mode ("grind mode")

- Cursor describes a **long-running** mode ("grind mode") that aligns on a plan first, then grinds until criteria are met—potentially for days [33].

7) "Meta-setup" is becoming its own benchmark (Karpathy)

- Andrej Karpathy says he has agents iterating on **nanochat** automatically: agents work on feature branches, try ideas, merge improvements, and iterate [34].
- In one snapshot he reports **110 changes in ~12 hours** reducing validation loss from **0.862415 → 0.858039** (d12 model) with no wall-clock penalty [35].
- He calls the real benchmark: "**what is the research org agent code that produces improvements on nanochat the fastest?**" [36].

8) Let the model improve the model (Hanson Wang's GPT-5.4 workflow)

- Hanson Wang says he asked **GPT-5.4-xhigh in Codex** to autonomously iterate on Codex's **own system prompt**; it ran **>17 hours**, executed

[28] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[29] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[30] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[31] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[32] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[33] Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
[34] post by @karpathy
[35] post by @karpathy
[36] post by @karpathy

200+ evals, wrote scripts to monitor eval progress, and pruned unpromising branches [37].

9) Skills need evals (not vibes): LangChain's skills benchmarking loop

- LangChain's Robert Xu outlines an evaluation pipeline: define tasks + define skills, run **with/without** skills, compare, iterate [38][39].
- Reported outcome (their tests): Claude Code completed tasks **82%** of the time **with** skills vs **9% without** skills [40].
- Practical detail: they stress **consistent clean environments** (they used a lightweight Docker scaffold) for reproducible agent tests [41][42].

10) Manual testing is still non-negotiable (and agents can help)

- Simon Willison: "Just because code passes tests doesn't mean it works as intended… Automated tests are no replacement for **manual testing**" [43].
- He recommends having agents **execute what they wrote** (e.g., Playwright for UI testing) instead of assuming correctness [44][45].
- For evidence, Willison's **Showboat** pattern records commands + outputs to discourage agents from writing what they *hoped* happened [46].

11) Security footgun: prompt-injected CI agents + cache poisoning (Cline)

- Cline ran an issue-triage workflow using `anthropics/claude-code-action@v1` on every newly opened GitHub issue with `--allowedTools "Bash,Read,Write,..."` [47][48].
- Because the workflow prompt included the untrusted **issue title**, an attacker could prompt-inject tool execution and use GitHub Actions cache behavior to poison shared caches and steal release secrets, leading to a compromised **`cline@2.3.0`** release (later retracted) [49][50][51].

[37] post by @hansonwng

[38] Evaluating Skills

[39] Evaluating Skills

[40] Evaluating Skills

[41] Evaluating Skills

[42] Evaluating Skills

[43] Agentic manual testing

[44] Agentic manual testing

[45] Agentic manual testing

[46] Agentic manual testing

[47] Clinejection — Compromising Cline's Production Releases just by Prompting an Issue Triager

[48] Clinejection — Compromising Cline's Production Releases just by Prompting an Issue Triager

[49] Clinejection — Compromising Cline's Production Releases just by Prompting an Issue Triager

[50] Clinejection — Compromising Cline's Production Releases just by Prompting an Issue Triager

[51] Clinejection — Compromising Cline's Production Releases just by Prompting an Issue

PEOPLE TO WATCH

- **Jonas Nelle + Samantha Whitmore (Cursor)** — unusually specific harness design details: test-first PRs, video review entrypoint, Slack-as-IDE, subagents, and long-running “grind mode” [52][53][54].
- **Michael Truell (Cursor)** — adoption signal: Automations running **thousands/day** internally, including “compute-intensive security review” and team memory [55].
- **Hanson Wang (OpenAI/Codex)** — concrete “agent improves agent” workflow (17h autonomous system-prompt iteration with 200+ evals) [56].
- **Andrej Karpathy** — framing shift: optimize the **agent org** (meta-setup) and measure “time-to-improvement” loops [57].
- **Simon Willison** — high-signal practical guidance across (1) **agentic manual testing** and (2) real-world **agent CI security failures** [58][59].
- **swyx** — pushes for better rigor + tooling around agent reliability, including an open-sourced **Claude compaction viewer** for diagnosing bad compactions [60] and a reminder that statistically meaningful SWE-bench comparisons can require **30–60x more compute** than cheap samples [61].

WATCH & LISTEN

1) Cursor Cloud Agents: test + video + remote desktop as the new review loop (02:23–05:33)

Hook: why video is the “entry point” for reviewing agent output, and how remote desktop/terminal access closes the loop on real verification.

Triager

[52] Cursor’s Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor

[53] Cursor’s Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor

[54] Cursor’s Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor

[55] post by @mntruell

[56] post by @hansonwng

[57] post by @karpathy

[58] Agentic manual testing

[59] Clinejection — Compromising Cline’s Production Releases just by Prompting an Issue Triager

[60] post by @swyx

[61] post by @swyx

Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor (2:23)

2) Slack as the collaboration surface for agents (20:57–23:26)

Hook: how agent threads + team follow-ups shift human work from "where does this if-statement go?" to product/UX decisions.

Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor (20:57)

PROJECTS & REPOS

- **Cursor Automations** (always-on agents): http://cursor.com/blog/automations [62]
- **OpenAI OSS Codex skills (curated list)**: https://github.com/openai/skills/tree/main/skills/.curate [63]
  - Example installer command shared by Peter Steinberger: `$skill-installer playwright-interactive` [64]
- **LangChain — skills benchmarking repo**: https://github.com/langchain-ai/skills-benchmarks/tree/main?ref=blog.langchain.com [65]
- **swyx — claude-compaction-viewer**: https://github.com/swyxio/claude-compaction-viewer/ [66]
- **Cloudflare — vinext (Next.js rewrite + migration skill)**: https://github.com/cloudflare/vinext?ref=blog.pragmaticengineer.com [67]

[62] post by @cursor_ai
[63] post by @steipete
[64] post by @steipete
[65] Evaluating Skills
[66] post by @swyx
[67] The Pulse: Cloudflare rewrites Next.js as AI rewrites commercial open source

    - The Pragmatic Engineer highlights the role of **comprehensive tests** in enabling AI-driven rewrites [68][69].
- **Clinejection attack write-up + cache poisoning tool**
    - Cacheract repo: https://github.com/adnanekhan/cacheract [70]
- **Simon Willison's agentic testing tools**
    - Showboat: https://github.com/simonw/showboat [71]
    - Rodney: https://github.com/simonw/rodney (prompting + screenshots) [72]

Editorial take: Today's theme is **throughput via autonomous + parallel agents**—and the tax you can't dodge is **verification (tests + manual evidence) and security boundaries** around what those agents are allowed to touch.

Sources

1. Cursor's Third Era: Cloud Agents — ft. Sam Whitmore, Jonas Nelle, Cursor
2. post by @OpenAI
3. post by @hansonwng
4. post by @OpenAIDevs
5. post by @embirico
6. post by @romainhuet
7. post by @cursor_ai
8. post by @jediahkatz
9. post by @jediahkatz
10. post by @jediahkatz
11. post by @cursor_ai
12. post by @mntruell
13. post by @jediahkatz
14. Qwen 3.5
15. The drama never ends…
16. post by @karpathy
17. post by @karpathy
18. Evaluating Skills
19. Agentic manual testing
20. Clinejection — Compromising Cline's Production Releases just by Prompting an Issue Triager

[68] The Pulse: Cloudflare rewrites Next.js as AI rewrites commercial open source

[69] The Pulse: Cloudflare rewrites Next.js as AI rewrites commercial open source

[70] Clinejection — Compromising Cline's Production Releases just by Prompting an Issue Triager

[71] Agentic manual testing

[72] Agentic manual testing

21. post by @swyx
22. post by @swyx
23. post by @steipete
24. The Pulse: Cloudflare rewrites Next.js as AI rewrites commercial open source