Shopify Shows Where Coding Agents Break Next

Coding Agents Alpha Tracker

2026-04-23

Shopify Shows Where Coding Agents Break Next

By Coding Agents Alpha Tracker • April 23, 2026

The strongest signal today came from Shopify: code generation is no longer the hard part—review loops, harness quality, and CI/CD are. Also in this brief: Slack-native agent launches from OpenAI and Cursor, a locally runnable Qwen3.6-27B setup, and the OpenClaw bug fix that changed agent behavior without touching prompts.

TOP SIGNAL

At Shopify, AI tool DAU is now near 100%, but CTO Mikhail Parakhin's practical takeaway is not run more agents. It is: use fewer agents, wire in critique loops, and spend heavily on PR review because the real bottleneck has moved to test failures, rollbacks, and CI/CD systems built for human-speed PRs [1].

> Too many agents in parallel that don't communicate with each other… are almost useless compared to just fewer agents [1]

That setup coincided with PR merge growth rising to 30% month-over-month from roughly 10%, while overall deploy time still improved despite slower reviews [1].

TOOLS & MODELS

- **Slack is becoming an agent control plane.** OpenAI launched **Workspace Agents** in ChatGPT: shared agents for complex tasks and long-running workflows across tools and teams, built on a cloud-hosted Codex harness with recurring tasks and Slack as one surface. Cursor shipped a similar Slack-side workflow: mention `@Cursor`, let it use thread + broader channel context, and review the PR it opens [2, 3, 4, 5].
- **Qwen3.6-27B**: the release claims flagship-level agentic coding performance beyond Qwen3.5-397B-A17B across major coding benchmarks. Simon Willison's local test ran the **55.6GB** full model or the **16.8GB** Q4

GGUF quant and saw roughly **24.7-25.6 tokens/sec** on SVG codegen tasks [6].

- **ADK 2.0 + Agent CLI**: Google's practical additions are clear enough to matter. **ADK 2.0** supports **Python, Go, TypeScript, and Java**, adds **graph-based workflows** for deterministic routing, and the new **Agent CLI** lets coding agents scaffold, deploy, evaluate, and add observability through natural-language commands [7].
- **Gemini Enterprise Agent Platform**: the production pieces to care about are gateway + agent identity + registry + anomaly detection, traceability, sandboxes, memory, and runtime support for agents that can keep state for up to **7 days** [7].
- **Claude Opus in OpenClaw:** Matthew Berman's heavy-use field report is still **Opus 4.6** for orchestration, mainly on personality and tool-calling. His warning on **Opus 4.7**: tokenizer changes can map the same input to roughly **1-1.3x** more tokens, and agentic runs can emit more thinking tokens too [8].

WORKFLOWS & TRICKS

- **Shopify-style critique loop.** 1) Let one strong model generate the code. 2) Hand the output or diff to a second model—ideally a different one—for critique. 3) Have the first agent revise. Parakhin says this beats non-communicating parallel agents on code quality, even though latency goes up [1].
- **Spend review budget, not just generation budget.** Shopify's review pattern is to use the largest models at PR time, have them take turns instead of swarming, and keep automated review strict. The claim is blunt: an hour of review is still cheaper than failed tests, hunting the bad PR, and rolling back deploys later [1].
- **Audit the harness before touching prompts.** OpenClaw had a bug where OpenAI-model setups silently fell back from **Codex harness** to **Pi harness**. After fixing auth and killing the silent fallback, the same prompts suddenly produced full agent loops, repo inspection, real edits, verification attempts, and continuity across heartbeats [9].
- **A local Qwen setup worth copying.** Simon Willison installed `llama.cpp` via Homebrew, then ran `llama-server` against `unsloth/Qwen3.6-27B-GGUF:Q4_K_M` with reasoning enabled and `preserve_thinking` turned on. That setup delivered local coding throughput in the mid-20 tokens/sec range on his SVG tests [6].
- **Thread-to-PR is now a real workflow.** Cursor's pattern is simple: mention `@Cursor` in Slack, let it read the thread and broader channels, watch progress stream back, then review the PR. OpenAI is pushing the same control-surface idea from the other direction with recurring tasks, tool hookups, and Slack-driven workspace agents [5, 3].

PEOPLE TO WATCH

- **Mikhail Parakhin** — rare source of production-scale numbers and anti-patterns: near-100% AI DAU, CLI agents growing faster than IDE tools, and a very clear view that Git/PR/CI/CD is now the bottleneck, not raw code generation [1].
- **Simon Willison** — still the cleanest operator for local-model testing. Today's Qwen3.6 note included the exact runtime stack, model choice, and real throughput numbers—not just benchmark screenshots [6].
- **@pashmerepat / OpenClaw contributors** — high signal because they proved a harness-layer bug can completely distort model behavior. If you're comparing agents, validate the plumbing first [9].
- **Alexander Embiricos** — worth following for where Codex is going next. His framing of workspace agents as cloud agents with their own identities, running a fully-powered Codex agent, is the clearest description of the product shift [4].
- **Adi Osmani, Dave Elliott, and Shubham Sabu** — the Google trio to watch if you care about production agent tooling beyond demos: today's drop included Agent CLI, graph workflows, long-running agents, and Agent Garden [7].

WATCH & LISTEN

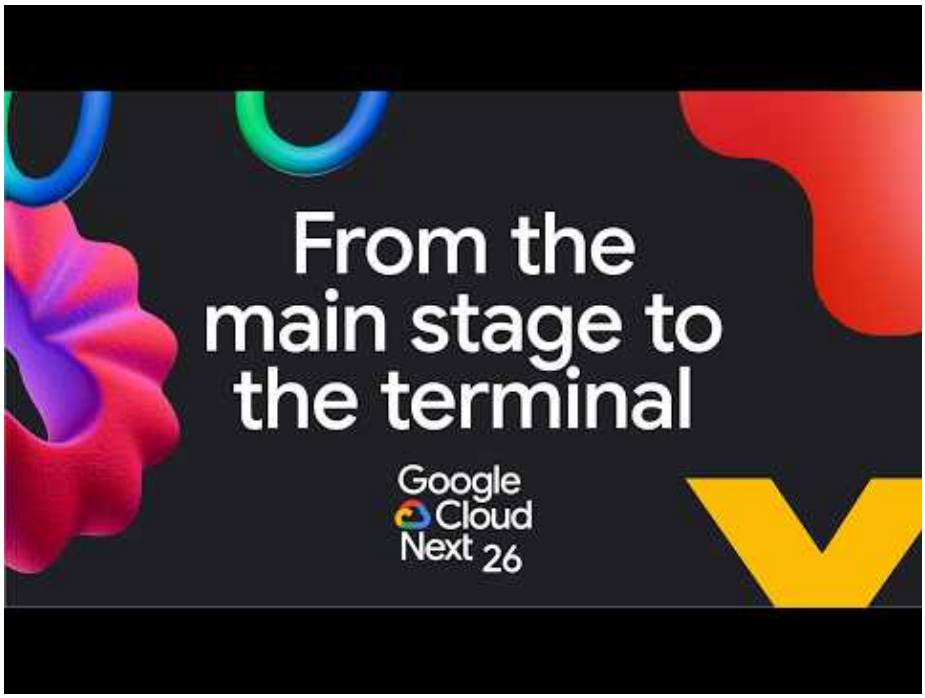
- **10:44-11:43 — Shopify on critique loops over swarms.** Best minute of the day if your default move is spawn more agents. Parakhin lays out the generate -> critique -> revise pattern and explains why different models debating beats parallel agents that do not coordinate [1].

CI/CD Breaks at AI Speed: Tangle, Graphite Stacks, Pro-Model PR Review — Mikhail Parakhin, Shopify (10:44)

- **15:23-16:07 — Why slower PR review can still ship faster.** Strong operations clip: longer model review latency is acceptable if it cuts failed tests, bad merges, and rollback churn downstream [1].

CI/CD Breaks at AI Speed: Tangle, Graphite Stacks, Pro-Model PR Review — Mikhail Parakhin, Shopify (14:44)

- **1:19:46-1:21:27 — ADK 2.0 graph workflows in plain English.** Useful if you need deterministic routing inside an agent system—for approvals, claims, or any workflow where you cannot let the model improvise every branch [7].



From the Next '26 main stage to the terminal (79:45)

PROJECTS & REPOS

- **codex** — Tibo says workspace agents are powered by Codex under the hood, using the same implementation open-sourced here [10].
- **Agent Garden** — the repo went live today and packages reusable templates for sequential, loop, parallel, human-in-the-loop, coordinator-dispatcher, and iterative-refinement workflows [7].
- **Qwen3.6-27B-GGUF:Q4_K_M** — the 16.8GB quantized build Simon used locally; the full **Qwen3.6-27B** model is **55.6GB** [6].
- **Droid Computers** — Factory AI opened access to persistent machines for remotely orchestrating Droids, each with its own filesystem, credentials, and configs. You can spin one up in Factory's cloud or turn your own machine into one; Ben Tossell says his Mac mini is already running as a Droid Computer [11, 12].
- **Cloud Run sandboxes** — secure, ephemeral, isolated sandboxes for executing agent-generated code, scripts, or Chromium from Cloud Run resources [13, 14].

Editorial take: the leverage is shifting away from spawning more agents and toward tighter review loops, cleaner harnesses, and better execution surfaces [1, 9, 3].

Sources

1. CI/CD Breaks at AI Speed: Tangle, Graphite Stacks, Pro-Model PR Review — Mikhail Parakhin, Shopify
2. X post by @OpenAI
3. X post by @gdb
4. X post by @embirico
5. X post by @cursor_ai
6. Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model
7. From the Next '26 main stage to the terminal
8. Wtf Anthropic
9. X post by @pashmerepat
10. X post by @thsottiaux
11. X post by @FactoryAI
12. X post by @bentossell
13. X post by @steren
14. X post by @simonw