# Claude Ship lets an agent handle PR review, fixes and policy-based approval

Coding Agents Alpha Tracker

2026-11-24

## Claude Ship lets an agent handle PR review, fixes and policy-based approval

*By Coding Agents Alpha Tracker • November 24, 2026*

Anthropic's Claude Ship applies agents to the work after a PR is opened. Also: what practitioners report from running Gemini 4 Neon, how Ladybird used agents to climb the web-platform-tests suite, GLM 5.4, and open-source agent patterns from LangChain.

### Claude Ship: agents now handle the work after the PR

Anthropic launched Claude Ship. It takes a PR through review, fixes and an approval decision based on a policy your team writes [1]. Boris Cherny's reason for it: models have written 100% of his code since Opus 4.5, but shipping PRs hasn't caught up. He says he uses Ship "all day every day" [2]. It has three parts:

- **Code Review (generally available):** reviews the PR against your codebase, checks its own findings adversarially, then classifies them and posts comments. You set the depth per repo: Quick, Standard or Deep [1].
- **PR Steward (beta):** responds to failing CI, merge conflicts and review comments. After each push it explains what it did and how to revert it. It does not approve or merge [1].
- **Claude Approvals (beta):** applies a policy file kept in the repo, like a "smarter CODEOWNERS." It approves routine changes or asks for a human in specific situations [1].

**How to try it:** upgrade the Claude Code CLI and type `/ship`. Claude pushes the change, opens the PR, reviews it and responds to feedback. Your permissions and branch protection still decide what merges [1]. On Team or Enterprise plans, admins can turn it on per GitHub repo. Cherny's advice is to start with one repo and run Approvals in observe mode until you trust it [3]. You pay per

token. Cost depends on how many PRs it touches, review depth and how much work PR Steward does, and admins can set spend limits [1].

Reactions focused on fit. Kent C. Dodds: "The big labs are eating your software factory product" [4]. One developer asked whether agent approvals meet regulations that require peer review [5]. Theo said that in most cases it's enough for the team to review and understand the intent of the change, or for someone to check a staging environment [6]. He is not a lawyer, so check your own compliance requirements.

## Gemini 4 Neon in real use: strong results, little visibility mid-run

DHH says Neon cut Omarchy's boot time in half and made big install optimizations. He calls it a "quiet cat": it shows little output while working, then delivers everything at the end [7]. He finds the harness "a little rough" and says it needs something like Claude's agent overview [8]. He ran Neon on 6 machines at once while optimizing for Intel, AMD and NVIDIA hardware, and never used up the session window. He isn't sure how long that generous usage will last [9].

Google released something that addresses his complaint the same day. Antigravity now has live progress checklists that track each step, and you can expand the full breakdown mid-run [10].

Ben Clavié puts Neon's capability around Opus 5.5. He says it is better designed to work with humans than any model he has seen, and it got him using Antigravity and T3 Code [11]. Theo says T3 Code's Antigravity bindings keep improving in the nightly build [12].

Fireship gives Neon's price as about $3 per million input tokens and $15 per million output tokens, with a 1M-token context [13]. It streams about 106 tokens per second, but Artificial Analysis counted 283M tokens on its test, more than 4x the median [13]. His suggested setup is the Antigravity orchestrator, with Neon directing a swarm of Hydrogen subagents [13].

## Ladybird alpha: a case study in agent-driven gains

Ladybird is now in alpha. Simon Willison's write-up shows how much of the progress came from agents. Most of a 75,650-subtest gain on web-platform-tests (WPT), which moved Ladybird ahead of Safari, came in five days. The team "pointed AI coding agents at WPT and asked them to fill in missing functionality." Many small fixes added up, about 30,000 of them in `/css` and 10,000 in `/editing` [14]. Earlier, Claude Code and Codex ported parts of LibJS to 25,000 lines of Rust in two weeks [14]. Which model versions they used isn't known [14].

The method can be reused: give agents a large conformance test suite and let them close the gaps. Theo used to think Ladybird was "the dumbest project

ever." He says agents "fundamentally changed my math" [15].

## GLM 5.4: first looped-transformer coding model, gated behind the Coding Plan

Z.ai's GLM 5.4 reruns a group of weight-shared layers several times. This makes the network deeper without adding parameters [16]. The weights are not open (yet?). For now you can only reach it with a Coding Plan subscription key, used against the OpenAI-compatible endpoint `https://api.z.ai/api/v1` [16]. Willison released llm-glm as a stopgap until `llm-zai` supports the model [16].

## Patterns worth studying

**Jev as a decision layer in agents.** In a LangChain interview, TypeSafe's co-founder says Jev can only return answers within the interface you specify, along with a confidence score. You set thresholds by how risky the decision is, calibrate before production, and send low-confidence cases to a smarter model [17]. The host, Sydney Runkle, named per-tool-call safety checks for coding-agent auto mode as a strong use. She had switched auto mode off because it was too slow, then back on [17].

**LangChain Content Studio (open source).** This design agent makes every graphic an HTML page and every video an animation timeline, so outputs stay editable. It is built on the Deep Agents harness and deploys with `uv run mda deploy`. Each thread gets its own Linux sandbox [18]. Some ideas you can reuse:

- Shared team memory sits alongside per-user memory [18].
- For long jobs, the agent drafts a storyboard before building, then builds scenes in parallel while you add notes. The final render waits for your approval [18].
- Recurring jobs become a skill [18].

## Smaller updates

- **Claude Code desktop:** Files, Changes, Terminal and other tools now open as tabs beside the session. You can view several at once or pop a tab out into its own window. Rolling out on Mac now, with Windows and Linux to follow [19].
- **Gemini API:** you can now send PDFs, Office files, audio and video and request only the pages, cells, segments or layers you need [20].
- **Willison** built datasette-charts "with a lot of help from Fable 5.2" [21] (try it [22]). Datasette Agent 0.7a1 adds side-panel navigation by fragment hash to support it [23].
- **OpenAI security:** OpenAI says GPT-6 Cyber helped validate 948 Linux kernel local-root bugs in one run and generated patches for review [24]. It

also released Shirudo with Cisco, an open-source runtime in which agents write eBPF shields so systems stay up while patches roll out [25].

---

**Sources**


1. X post by @addyosmani
2. X post by @bcherny
3. X post by @bcherny
4. X post by @kentcdodds
5. X post by @alexandereardon
6. X post by @theo
7. X post by @dhh
8. X post by @dhh
9. X post by @dhh
10. X post by @antigravity
11. X post by @bclavie
12. X post by @theo
13. Opus just lost to the last company you'd expect…
14. Ladybird's first alpha release
15. X post by @theo
16. GLM-5.4: Toward Generalized Coding with Our First Loop Transformer
17. Jev Explained: How TypeSafe's System One Model Fits into Your Agents
18. We Open Sourced Our Design Brain
19. X post by @ClaudeDevs
20. X post by @OfficialLoganK
21. X post by @simonw
22. X post by @simonw
23. datasette-agent 0.7a1
24. X post by @OpenAI
25. X post by @OpenAI