

# The Merge, Not the Diff, Is the New Agent Bottleneck

Coding Agents Alpha Tracker

2026-09-13

## The Merge, Not the Diff, Is the New Agent Bottleneck

*By Coding Agents Alpha Tracker • September 13, 2026*

Omarchy's 2,200-PR backlog and Astra's reliability reset show where coding-agent leverage is moving: agents prepare and verify work, while humans own trusted merge units and product judgment.

Coverage is incomplete: some monitored sources or documents could not be processed. This brief covers the available verified material.

Coverage is incomplete: this brief reflects the available verified material.

### TOP SIGNAL

**Omarchy is treating agent autonomy as a triage-and-merge problem.** DHH says the project is 2,200 PRs deep and receiving nearly 100 new submissions per day; agents will help, but humans are still needed. [1] The proposed unit of work is a trusted cluster: triage consolidates duplicates and prepares finished submissions, while agents and people review batches before whole tranches are merged—and user-experience decisions remain human-owned. [2, 3]

**The pattern to steal:** make agents produce reviewable merge units, not an ever-growing pile of individually plausible diffs.

### TRY THIS

- **Encode the bar instead of repeating it in chat (Addy Osmani).** Start with the outcome, definition of “done,” non-touch constraints, and whether reuse or refactoring is simpler. Then give the agent the exact build, test, and lint commands; turn recurring review rejects into reusable checks such as `/verify`, `e2e`, and schema checks; run them before opening

the PR and use `/code-review`. When the agent misses, write the lesson into `CLAUDE.md` or a skill rather than quietly fixing it by hand. [4]

- **Batch PR triage, then merge tranches.** For a noisy repository, have an agent consolidate duplicates and cluster related fixes. Require manual and agent review plus serious QA effort for each cluster; keep product and UX acceptance with a human. This is DHH’s response to Omarchy’s backlog, not a claim that agents can safely auto-merge it. [1, 2, 3]
- **Provision the environment before multiplying agents.** Armin Ronacher found a controlled machine reachable over SSH more useful for day-to-day engineering than cloud environments where PostgreSQL was absent and database tests were mocked. Seven parallel agents can mean seven databases, Kafka clusters, and other service instances; even cached environment bootstrap was estimated at roughly seven minutes. Start with one faithful environment, then fan out only when each worker has isolated dependencies. [5]
- **Use a real artifact request as a capability probe.** Simon Willison’s prompt was: I live at <my address>. Figure out 5K and 10K running routes from me that loop from my house. Use OSM data. GPT-6 Astra ran for 27 minutes and returned an embedded visualization plus GPX and GeoJSON files. Inspect the artifacts—not just the prose—and, if useful, check whether the generated map is reproducible: Willison traced his result to a `visualize` skill emitting D3-based HTML. [6, 7]

## WHAT SHIPPED

- **Astra’s reliability reset fully propagated across Codex and ChatGPT Work.** The product-team update says the reset addressed skills that over-triggered or prevented self-checks, context management that caused early stops or stale replies, and misconfigured engines that degraded quality. [8, 9]
- **DwarfStar support for DeepSeek Flash v4.1 is becoming a reproducible local-agent path.** In Salvatore Sanfilippo’s demo, a 2-bit quantized model ran 50/50 across two MacBook M5 Max machines with 128 GB each, reaching roughly 29–30 tokens/s decoding and 700–800 tokens/s prefill; he notes that the quantization had not yet been benchmarked. [10] The more interesting control surface is a smooth 0–100 thinking-strength setting that can be raised for a hard step and lowered afterward, combined with image inspection: the model rendered a Mandelbrot program, converted output to PNG, checked colors and zoom levels, and iterated. [10]
- **The strongest desktop-agent setup is still hardware- and context-dependent.** Riley Brown says he is experimenting with GrokBot, GPT

Work, Meta Muse, and Claude Cowork without committing to a cloud setup. His known-good pattern is a remotely controlled 48 GB Mac running Codex + Astra, with browser accounts already signed in, access to an external video drive, Notion, and Google Drive; he reports that it can control DaVinci Resolve and Premiere and is more predictable than Open-Claw, at the cost of heavy token use. [11]

## GO DEEPER

- **Boris Cherny — practical agent workflow.** Start with codebase Q&A—how code is used, Git history, and linked issues—before editing. For a large change, ask **Before you write code, make a plan**, approve it, then give the agent a verifier such as unit tests, Puppeteer screenshots, or an iOS simulator so it can iterate. Cherny also recommends short shared project context, checked-in MCP configuration, and treating the SDK as a JSON-producing Unix utility for CI and incident-response pipelines. [12] He reports that about 80% of Anthropic’s technical staff use the tool daily, including researchers editing and running notebooks. [12]



*Claude Code’s Creator Just Revealed 40+ Features Nobody Talks About (7:42)*

- **Armin Ronacher — Pi Agent dev workflow.** The useful section is the case for a minimal, shell-first harness: Bash and `rg` can pipeline discovery and transformations without dragging every intermediate result into context. The same discussion is a reality check on cloud-agent enthusiasm: local machines win when stateful services and private data matter,

while Ronacher says he sees much clearer gains in side projects than in enterprise rollouts and does not yet see strong outcome evidence behind higher commit volume. [5] Resources: pi.dev and arendel.com.



*Pi Agent dev reveals his Agentic Engineering Workflow (24:02)*

- **Salvatore Sanfilippo — DeepSeek Flash v4.1 + DwarfStar.** Skip the model-size discussion and watch the vision-in-the-loop segment: the agent generates a terminal program, inspects a PNG of its own output, tests multiple regions and zoom levels, and uses visual feedback to improve



the result. [10]

*Proviamo DeepSeek v4.1 Flash a 2 bit con DwarfStar (SUB ITA) (7:44)*

**Editorial take:** The winning agent loop is not “more autonomy”; it is explicit constraints, feedback-rich outputs, environment fidelity, and a human-owned merge bar. [4, 5, 2]

## Sources

1. X post by @dhh
2. X post by @dhh
3. X post by @dhh
4. X post by @addyosmani
5. Pi Agent dev reveals his Agentic Engineering Workflow
6. Generating running routes with GPT-6 Astra and ChatGPT Work
7. X post by @simonw
8. X post by @reach\_yb
9. X post by @thsottiaux
10. Proviamo DeepSeek v4.1 Flash a 2 bit con DwarfStar (SUB ITA)
11. X post by @rileybrown
12. Claude Code’s Creator Just Revealed 40+ Features Nobody Talks About