Ronacher and DHH: replace line-by-line review with test suites, benchmarks and agent-written guides

Coding Agents Alpha Tracker

2026-11-28

Ronacher and DHH: replace line-by-line review with test suites, benchmarks and agent-written guides

By Coding Agents Alpha Tracker • November 28, 2026

Armin Ronacher's DevFest keynote and a DHH interview both describe verification setups that let agents take on whole pieces of work. Also: DHH makes Gemini 4 Neon his default for systems work, the PR as a unit of work, and v2 for agentic video editing.

Ronacher: give the agent a test suite, then read its guides instead of its diffs

Armin Ronacher gave the opening keynote at DevFest Vienna. The pattern he says has run through his last six months: if you have a binary, measurable reference, an agent can rebuild software from scratch against it. Over Christmas he ported a templating engine he wrote in Rust to Go. He gave the agent only golden snapshot tests (an input and its expected output) and didn't require it to match the original implementation. "For the most part, I think this finished within two days" [1].

Other ideas from the talk you can reuse:

- **Agent-written guides as review.** An agent maintained a kids' guide to a game he was building, linked to the code so the two stayed in sync. The guide surfaced a bug: he wanted seeded worlds to be fully deterministic, but the order in which blocks are discovered was not. "There's no chance you would have caught this from the code" [1]. He now has the agent write similar "lectures" for his JavaScript engine instead of reviewing its

code changes, because the lectures call out what was unexpected [1].

- **Throwaway code goes through JavaScript.** Most code an agent writes is single-use code for calling tools, benchmarks and validation. The harness has no control over what Bash runs, so Pi adds a "code mode" that runs JavaScript. Ronacher picked JavaScript because models are trained on it and, unlike most dynamic languages, it has a spec [1].
- **Long tasks need durable infrastructure.** A task that takes 15 minutes is just a bash loop. Tasks that run for a week and spawn many subagents have to survive being stopped and moved, then continue where they left off [1].

He says about 90% of his work over the past month has been "idea comes in, code comes out." In his view the bottleneck is no longer code review but human communication within a company [1].

DHH: verify instead of reviewing every line, and give every app a CLI

In an Open AGI Forum interview, DHH said he has already written many applications without looking at a single line of their code. For work you can judge by its output, he argues against human line-by-line review. His alternatives are benchmarks, automated tests, fuzz testing, and "adversarial agent review where you have an agent from a different maker review the work." The human's job is to specify the software, judge whether it's good enough, and decide when it's done [2]. Optimization works especially well when every agent action returns a clear signal, such as whether the program got faster [2].

His data point: Fable translated TTFX, a Python graphics library used for Omarchy screensavers, into a pixel-perfect Rust version in 42 minutes. The Rust was 10× faster, started 40× faster, and shipped as a single 3 MB executable with no dependencies. Optimization loops took it to 30× faster. An assembler rewrite took a couple of hours and reached 100× [2].

Other habits he described:

- Every new application he makes gets "a great CLI," because agents work best through CLIs and computer use is slower [2].
- Disposable UIs: while reviewing open PRs in his Omarchy repository, he had Claude build a Tinder-style interface where he swipes to either dive into or merge each PR. He may just throw it away [2].
- His prompts are now deliberately vague. He gets a working version first and iterates, because he won't know what he wants until he sees it [2].

Gemini 4 Neon is DHH's default for systems work

After a week of use, DHH made Neon his "default driver for anything to do with systems work and optimization" and says it's the first model he installs on a new machine [3]. The caveats: it's "a fair bit slower" than other frontier models, so

he doesn't use it for simple tasks, and it "can't hold a candle to Opus/Fable on design" [4]. He reports it fixed everything on his ZenBook A16 X2E, including deep sleep [5]. He also amplified an Omarchy user's report: after Grok, Fable and Astra couldn't explain three months of slow reboots, Neon traced them to a failing NVMe SSD in a couple of minutes [6]. ThePrimeagen, meanwhile, asked whether Neon has been "nerfed" [7].

The PR as a unit of work

Sunil Pai (@threepointone) says the economics of software factories come down to two questions: how much a PR is worth to you, and how many you want to merge each day [8]. Kent C. Dodds replied that PRs "should not be the optimized primitive." His agents often want to split work into several PRs, but he prefers one PR per unit of work, unless there's a good reason to split, such as a migration that has to happen over time [9].

Tools and projects

- **Google Antigravity:** paste a GitHub PR link into chat to get an inline "PR pill." Hover over it to see status, branches and comments, and find every linked PR in the Overview menu [10].
- **v2** (Riley Brown): a free, open-source "GitHub for agentic video editing." It stores every version of a video with the prompts, files and comments behind it, so agents learn what you like. Anyone can fork a public version with one prompt. v2 is not an editor; it works with Tesseract, Hyperframes, Remotion and others. Brown says his team has used it internally for a month and all of its editing is now done by AI. CLI: agentnativeinc/v2 [11].
- **deser:** Ronacher suggests spending some agent tokens on investigating deser as an alternative to serde for Rust. He doesn't care whether you adopt it; he wants to know if the design has a future [12].
- **RL task design** (Gabe Greenberg, G2i, on Latent Space): RL training tasks should be "a few steps away from what the model can do today." Tasks that are too hard become evals, and tasks that are too easy are worthless [13].

Sources

1. Agentic Insanity | Armin Ronacher | DevFest Vienna 2026 Opening Keynote
2. Jianzhong Li in Conversation with DHH:10 Key Topics in the Agentic Al Ecosystem Open AGI Forum
3. X post by @dhh
4. X post by @dhh
5. X post by @dhh

6. X post by @Gardnmi
7. X post by @ThePrimeagen
8. X post by @threepointone
9. X post by @kentcdodds
10. X post by @antigravity
11. X post by @rileybrown
12. X post by @mitsuhiko
13. Inside the AI Training Data Market — Gabe Greenberg, G2i