Agents Need Measurement Loops—and a Conductor

Coding Agents Alpha Tracker

2026-08-05

Agents Need Measurement Loops—and a Conductor

By Coding Agents Alpha Tracker • August 5, 2026

Theo's T3 Code debugging postmortem and Kent C. Dodds' conductor point to the same practical frontier: agent systems that measure, hand off, and recover instead of merely generating patches.

TOP SIGNAL

Theo's T3 Code performance postmortem is a useful boundary for agentic debugging: a vague Codex request produced a confident diagnosis and a 10,000+ line PR that changed nothing. The breakthrough was to stop asking for a fix and have the agent build a console-driven toggle harness; testing the hypotheses isolated an infinite sidebar opacity animation. The human supplied the hypotheses and validation—the agents were valuable as fast codebase searchers and diagnostic-tool builders, not autonomous diagnosticians. [1]

TRY THIS

- **Turn bug reports into experiments.** Isolate the target in a one-tab browser and use Task Manager: Theo notes that browser tooling is weak for CSS/compositor work, while DevTools changes performance characteristics. Ask the agent to build a console-pasteable toggle for suspected effects, apply all, reset, then flip one feature at a time; separate transitions from animations before editing. Theo got the GPU process down to 3% or less with the toggles applied, back above 25% after reset, and eventually traced the worst offender to the sidebar terminal icon's pulse. [1]
- **Put a conductor over the fleet.** Kent C. Dodds' pattern is one supervisor spawning isolated Cursor Cloud Agents, using the Kody Koala

MCP for handoffs, creating a PR when a worker stalls, receiving completion messages, and sending a Discord summary at the end. Make each worker's environment, handoff state, and completion message explicit; the control plane is more valuable than another giant prompt. [2]

- **Use near-free models as sidecars.** Theo says Luna became effectively free after an 80% cost reduction and is using it for T3 Code title generation; he wants it on every prompt for descriptions, feedback, and statuses. Route low-risk metadata and auxiliary outputs there, but measure whether the extra calls improve the workflow before letting cheap inference become unbounded background work. [3]
- **Normalize shared skills with a compatibility shim.** DHH reports that Claude Code still does not natively scan `~/.agents/skills`; his workaround is a symlink, despite the Claude docs acknowledging the pattern. Keep skills in one canonical tree, symlink where needed, and add a fresh-machine discovery check to your agent setup. [4, 5]

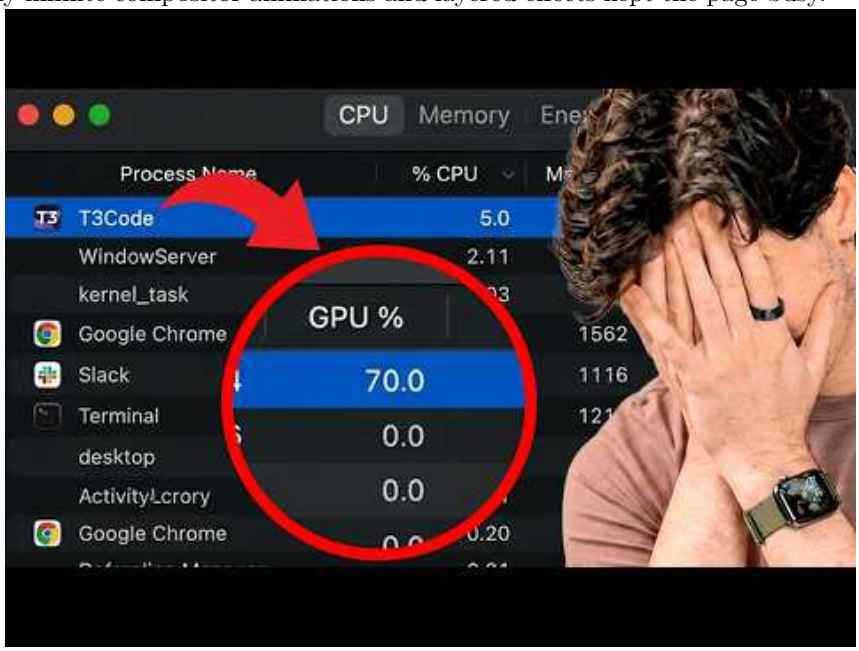
WHAT SHIPPED

- **LLM 0.32:** Simon Willison's major CLI/library release sends reasoning traces to stderr (`-R/--hide-reasoning` suppresses them), adds the GPT-5.6 family with GPT-5.6 Luna as the default, and supports server-side Code Interpreter, WebSearch, WebFetch, CodeExecution, and MCP tools. The new `llm openai endpoint` command can run one-off prompts against any OpenAI-compatible endpoint—including a local LM Studio model—without logging them. [6] It also adds typed `stream_events()` for mixed reasoning/text/tool outputs and tool-chain pause/resume from stored history: primitives worth copying into any coding-agent loop that needs human gates and durable state. [6]
- **OpenWiki 0.3:** A full codebase-wiki prompt rewrite reports a 28.57% relative success increase (35% → 45% at `n=2`), 14% fewer tokens, and 26% fewer tool calls per successful task. Install with `npm install -g openwiki@0.3.0`; treat the numbers as an early, self-reported signal and rerun the eval on your own repositories. [7]
- **Model routers are becoming a coding-agent layer.** Not Diamond announced routing across gateways and harnesses, including Claude Code, claiming 20–65% lower cost without a quality hit. Mckay Wrigley sees the larger opportunity in blending "jagged" models into smoother behavior, describes router engineering as a third layer after model and harness engineering, and says DeepSeek V4 Flash was cheap enough to offload roughly a half-dozen tasks from his Fable 5 workflow. The cost claim is vendor-reported; the actionable test is per-task routing and blending, not headline token price. [8, 9, 10, 11]
- **Resilience and fallback primitives:** LangChain says Deep Agents,

LangGraph, and LangChain can retry interrupted work, follow a safe recovery path, resume from saved state, and fall back to alternate models. Its Gateway announcement adds fallback rules across models and hosts when a provider fails or rate-limits. This is the right production direction: recovery should be part of the agent runtime, not a human restarting a dead run. [12, 13]

GO DEEPER

- **Video — Theo's T3 Code performance postmortem.** Focus on the diagnostic-harness segment: the agent becomes useful when the engineer turns competing theories into measurable toggles, then the video explains why infinite compositor animations and layered effects kept the page busy.



[1] *Fable Broke My App and Couldn't Fix It (11:54)*

- **Repo — OpenWiki.** Study the prompt rewrite as an example of improving an agent by changing its codebase-understanding instructions rather than swapping models; reproduce the success, token, and tool-call measurements before trusting the tiny `n=2` sample. [7]
- **Agent framework — llm-coding-agent.** LLM 0.32's lower-level work was driven by Datasette Agent and llm-coding-agent; inspect the combination of model/tool mixing, structured streaming, human approval, and resume-from-history rather than treating an agent as a single prompt wrapper. [6]

Editorial take: The frontier is shifting from prompt quality to control-plane

quality: humans design the measurement, agents build the probes and patches, and supervisors carry state between workers. [1, 2]

Sources

1. Fable Broke My App and Couldn't Fix It
2. X post by @kentcdodds
3. X post by @theo
4. X post by @dhh
5. X post by @dhh
6. New release of LLM adds support for reasoning traces, OpenAI Responses, server-side tools, and smarter logging
7. X post by @BraceSproul
8. X post by @tomas_hk
9. X post by @mckaywrigley
10. X post by @mckaywrigley
11. X post by @mckaywrigley
12. X post by @LangChain
13. X post by @LangChain