

Coding Agents Are Moving From Prompts to Control Graphs

Coding Agents Alpha Tracker

2026-08-26

Coding Agents Are Moving From Prompts to Control Graphs

By Coding Agents Alpha Tracker • August 26, 2026

The most useful signal today is operational: explicit state, verifier gates, and trace-driven evals are the practical path beyond one-shot prompting. New T3 Code and OpenWiki releases make that shift concrete.

TOP SIGNAL

AI Jason's practical thesis is that repeatable coding work should be designed as a control graph, not left as a chat prompt. His team models workflows as nodes, edges, and carried state; its change-shipping graph runs setup/implementation → verification → simplification/PR only when verification passes, with structured outputs passed between agent sessions. [1] The design keeps the human one layer above execution, but only after a separate verifier and deterministic scripts take over self-checking, data fetching, server startup, and tests. [1]

TRY THIS

- **Map one recurring loop before adding more agents.** Write its SOP; label each node (action), edge (next transition), and state (data carried forward); specify whether the trigger is time-, goal-, or event-based and which conditions require a human. Keep current state in a Markdown status file and append each run to a log. Use scripts for complex fetching, server startup, evaluation, and end-to-end tests. [1]
- **Make verification a hard gate.** For a scoped change, run `setup/implement → verify → simplify/PR`; do not enter the last phase unless verification passes. Give every agent node an explicit model

and output schema, then build the next prompt from the prior node's structured output. [1]

- **Turn traces into an eval factory.** Load the eval-engineering skill. Use: `Use the eval-engineering skill, the traces from {LangSmith project}, and the {current repository} to help me create an eval Task for {agent}.` [2] Review the generated world-spec skill, validate it in a new thread, iterate until reliable, then ask for 10 new task specs from recent traces and instantiate them. [2] Run tasks with real agents and multiple model tiers to expose leaky environments and reward hacks; keep human review for domain alignment and difficulty calibration. [2]
- **Protect the prompt cache.** Pick the model and effort level at session start. Addy Osmani says switching later forces a full uncached reread because the KV cache is tied to model weights; effort level and fast mode have the same behavior. Switch only in the first few turns if you must. [3]

WHAT SHIPPED

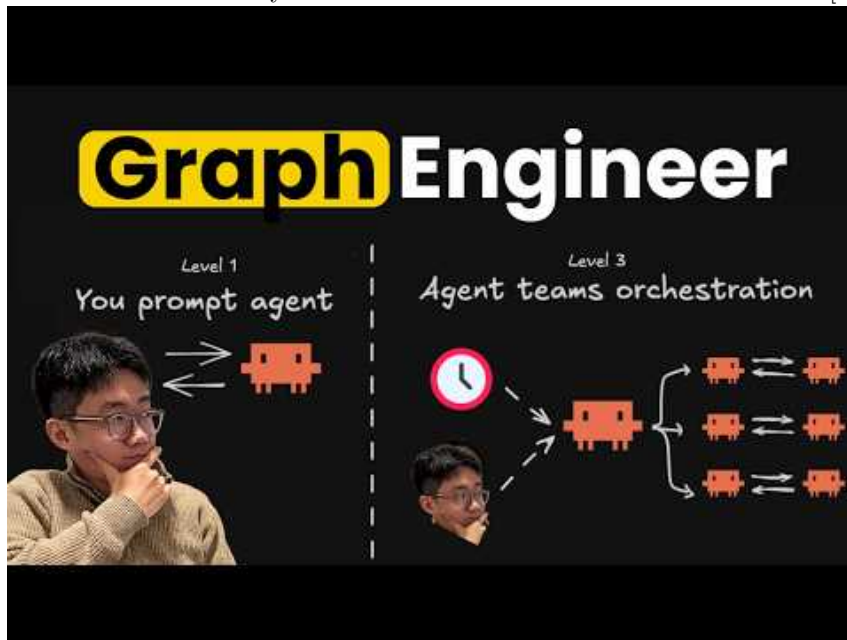
- **T3 Code — 380+ changes.** Theo's release adds in-app PR reviews, `npx t3 triage`, `Mac npx t3 connect`, a 93.8%+ improvement in long-thread rendering, an 80%+ reduction in thread data transferred and stored, one-click VS Code-over-SSH projects, and an overhauled terminal. It also improves Codex/Claude/OpenCode skill discovery and surfaces Claude compaction recommendations plus Codex MCP/ComputerUse permission requests. Nightly builds expose the features before stable, with Theo's caveat that they are “mostly” done. [4]
- **OpenWiki 0.4.0 — evidence-backed memory.** It records each factual claim with supporting code and an evidence version; changed evidence flags the claim stale, which persists until re-verification. Staleness checking runs deterministically before the agent and unresolved claims stay flagged. [5] In the article's own replay, stale claims fell from 80 to 9 and hallucinated claims from 15 to 0; upgrade with `npm install -g openwiki@latest` and `openwiki --init`. [5]
- **Rails' lemans — open-source agent eval harness.** The CLI-first harness supports Daytona sandboxes and a local Docker backend. It hides `verification_test.rb` and restores `test/`, `bin/`, and `config/environments/test.rb` before grading, keeping the agent from changing the graded surfaces. [6] Its new-model report is a useful task-specific comparison: Terra scored 49/63 at a 182-second median; open-weight Qwen 3.8-27B scored 48/63 but took 27 minutes and had 7.9% Rails API recall; Sonnet 5 scored 44/63 despite higher API recall than Opus 4.8. [6]
- **Event-triggered ChatGPT Work tasks.** Plus and Pro users can now

run tasks when GitHub, Slack, or Gmail changes rather than only on a fixed schedule; the Free rollout supports up to three scheduled tasks, and tasks can be shared for customization. [7]

- **ChatGPT Sites — useful but still a demo.** Riley Brown used @sites in ChatGPT Work to generate a hosted app with a database, authentication, storage, and custom-domain support; his prompt included team sign-in, video storage, and an API-backed skill. The run took 25 minutes 8 seconds, after which he used browser annotations to request UI fixes. He says it had not yet been sent to his team, and the video was OpenAI-sponsored; use environment variables rather than pasting API keys. [8]

GO DEEPER

- **AI Jason — “I don’t prompt agents anymore...”** Watch the concrete design: deterministic heuristics filter bad designs, parallel screenshot agents evaluate the survivors against user requests, shared schemas standardize outputs, and a main agent ranks the report; his code-shipping workflow then uses dynamic workflow with structured handoffs. [1]



I don't prompt agents anymore... (16:04)

- **DHH — “Is Linux the perfect operating system?”** The crash-to-PR segment is a useful risk boundary: diagnose → report → permissioned GitHub submission → maintainer-agent PR. He then shows a low-criticality library being patched, released, and tagged as 0.31, while explicitly saying he would not automate everything. [9]



Is Linux the perfect operating system? - Omarchy interview with DHH (14:28)

- **Study the guardrails in lemans and OpenWiki.** One isolates verification so evals are harder to game; the other makes project memory stale when its evidence changes. Both are better starting points than adding another prompt or memory layer. [6, 5]

Editorial take: Reliable coding-agent leverage is becoming an operating loop—explicit state, a separate verifier, and bounded permissions—not a more elaborate prompt. [1, 9]

Sources

1. I don't prompt agents anymore...
2. X article by @Vtrivedy10
3. X post by @addyosmani
4. X post by @theo
5. X article by @colifran__
6. Agents on Rails: lemans goes open source
7. X post by @ChatGPT
8. NEW ChatGPT Sites Just Changed Everything (Builds Anything)
9. Is Linux the perfect operating system? - Omarchy interview with DHH