The Harness Beats the Model: Self-Optimizing Inference, Compaction, and Cross-Model Checking

Coding Agents Alpha Tracker

2026-07-30

The Harness Beats the Model: Self-Optimizing Inference, Compaction, and Cross-Model Checking

By Coding Agents Alpha Tracker • July 30, 2026

Three independent signals — GPT-5.6 Sol optimizing its own serving kernels, ARC-AGI-3 SoTA via compaction settings, and Tim Dettmers' harness outperforming newer Opus models — converge on the harness layer as the real frontier for coding agents.

TOP SIGNAL

The harness layer is where the gains are — not the model layer. Three independent signals converge today. Tim Dettmers (creator of bitsandbytes, QLoRA) reports that Opus 4.6 combined with his custom harness significantly outperforms Opus 4.8, and Opus 5 is "not a large improvement either" — then adds: "There are such easy gains to be made with the harness. Really not sure what OpenAI and Anthropic doing." He plans to open-source a harness optimized for open-weight models handling tasks exceeding 10M tokens [1, 2]. Meanwhile, OpenAI applied GPT-5.6 Sol to optimize its own serving infrastructure: Codex analyzed production traffic, rewrote GPU kernels, and ran hundreds of experiments on its speculative-decoding model, cutting end-to-end serving costs by 20% and improving token-generation efficiency by 15%+ [3, 4]. And GPT-5.6 Sol reached SoTA on ARC-AGI-3 not through a model upgrade but through two harness-level setting changes: allowing the model to reason across multiple context windows using a "canonical compaction implementation" [5]. Peter Steinberger mocked Anthropic's earlier "victory" tweet about Opus 5's 3x ARC-AGI-3 lead now that two settings closed the gap [6]. FactoryAI CTO Eno Reyes reinforces the point from the demand side: "If I don't

know what model I'm using, I say it's great. If I know it's the frontier model, my bias kicks in" — arguing much of the "model wars" narrative is marketing [7].

TRY THIS

- **Cross-model checking: have one agent verify another's work.** DHH: "I've found it incredibly useful to have Opus 5 check the work of Sol or vice versa" [8]. Riley Brown applies the same pattern universally: "For every task now I tag in another agent to check the other agents work" — and calls it "harness agnostic software" [9]. No orchestration framework needed — just a second opinion from a different model. This is the simplest multi-agent pattern that directly addresses Opus 5's 50% hallucination rate [10].
- **Orchestrate agent teams in Buzz with a "take the lead" prompt.** Buzz (free, open-source, by Jack Dorsey) is a Slack-like interface where agents are channel members, not add-ons. The core delegation prompt that works: "One of you take the lead, figure this out" — agents spontaneously self-organize, with one taking the lead and consulting others [11]. Pin agents to specific models for cost routing: Fable for complex planning, Sonnet for simple reviews and summaries [11]. To add any OpenRouter model (e.g., Meta's Muse), create a Buzz Agent, set LLM provider to "OpenAI compatible," paste your OpenRouter API key, set base URL to `https://openrouter.ai/api/v1`, and set thinking effort to "Inherit agent defaults" [11]. Control parallelism per agent under Advanced settings (1–20 concurrent tasks) to stretch a single subscription [11].
- **Build a management agent that triages your inbox every 3 hours.** Riley Brown's #1 Buzz use case: a Codex-powered agent in a dedicated management channel that reads email, Slack, and texts at 9am, 12pm, 3pm, and 6pm, then outputs a ranked action list. He replies directly in the channel to execute ("write this email back to this person") [11]. The agent inherits all of Codex's existing skills but stays focused via a narrow one-paragraph system prompt [11]. The same pattern works outside Buzz — 37signals runs "Agent Marie" via the Basecamp CLI, posting cards, comments, reports, bug fixes, and PRs as if a human team member, with no special AI features in Basecamp itself [12, 13].
- **Use the loop-engineering pattern: cron + cheap pre-stage + agent only when needed.** Jason Zhou's autonomous Reddit karma loop grew an account from -4 to 95 in 7 days using an architecture that generalizes beyond Reddit. A fixed cron fires every half hour (reliability); a cheap deterministic workflow checks guardrails and rolls dice (randomness + cost control); the expensive agent only wakes on slots that pass [14]. Ground the agent in a personal wiki — it can only cite positions you've documented, never reconstruct opinions from memory — to avoid generic

slop [14]. Every Sunday, the loop self-evaluates: re-fetches all comments, computes karma delta by subreddit and topic angle, then rewrites its own strategy notes from data [14].

WHAT SHIPPED

- **Buzz** — Free, open-source Slack-like platform by Jack Dorsey for multi-agent orchestration. Connects to existing Claude Code, Codex, Cursor, and Grok Build subscriptions via the Agent Connect Protocol; agents run as CLI harnesses in the terminal [11]. Riley Brown calls it "a really important form factor for AI agents" that "requires no technical ability" [11].
- **T3 Code on iOS and Android** — Theo's free, open-source app for remotely controlling Claude and Codex. Run `npx t3 connect`, install the app, control agents from your phone. Hit 150,000 users and 30,000 weekly actives within hours of launch [15, 16, 17].
- **Cursor on iPad** — All the power of Cursor on iPhone, with more room to work with agents [18].
- **OpenWiki + LangSmith tracing** — The open-source codebase wiki generator now connects to LangSmith traces to analyze where coding agents (Claude Code, Codex) lack context or get stuck, then generates more targeted wikis. Try locally: `npm install -g openwiki` then `openwiki --init`. Repo: github.com/langchain-ai/openwiki [19, 20].
- **LangChain Academy: Autonomous Agent Improvement with LangSmith Engine** — New course on identifying and prioritizing issues from traces, drafting fixes, and proposing evals to prevent regressions [21].
- **GPT-5.6 Sol self-optimization** — After deployment, OpenAI applied Sol to improve its own inference: 20% lower serving costs from GPU kernel improvements, 15%+ better token-generation efficiency from improved speculative decoding [4]. Simon Willison: "Presumably that's billions of dollars a month in savings at this point?" [22].
- **Opus 5 analysis (Fireship)** — 1M token context window, 128K output tokens, five thinking levels (Low through Max). Promises near-Fable intelligence at half the price and verifies its own work without human intervention — but hallucination rate jumped 14 percentage points to 50%, and the model is "more neurotic," producing longer responses and sometimes doing more than asked [10].
- **Geoffrey Huntley: use Temporal.io for agent orchestration, not n8n.** Dismisses n8n as built by people who "have never worked in corporate" and don't know service buses are a solved problem. Recommends Temporal.io with a job invoking an agent as a process — "don't make the

entire service bus non-deterministic" [23]. Separately argues "software factories" are real but uncracked: the bottleneck is systems engineering (sandboxing, monorepo, reproducible builds, CI/CD, identity/secret management), not tokens [24, 25].

- **Similarweb's Deep Research agent eval framework** — Four methods wired to LangSmith traces: deterministic checks for tool calls, rubric-scored LLM judges for quality, faithfulness checks against retrieved data, and A/B comparisons against a saved baseline [26].
- **OpenAI researcher access** — Free frontier model access for scientists, starting with 10,000 researchers and expanding to 100,000 through 2027 [27].

GO DEEPER

- **Buzz: Building an Agent Team (Complete Guide)** — Riley Brown's full walkthrough with guest Vinnie (@hot_town). Covers the "take the lead" delegation pattern, management channel setup, OpenRouter integration, and the Task Checker workaround for buggy recurring workflows. The closing segment on Riley's management agent is the most concrete production agent setup documented this week.

Claude Code + Codex Can FINALLY Work Together (Buzz AI) (51:50)

- **Fireship: Did Anthropic just kill the indie hacker?** — Opus 5's

near-Fable intelligence at half the price means "execution costs $20 per month and anyone can build their own personal software instead of pay some random SaaS product." The indie hacker moat was coding itself — and that moat is gone.

Did Anthropic just kill the indie hacker…? (2:45)

- **ThePrimeTime: No Slop Allowed (Codeberg bans AI code)** — Codeberg's Terms of Use now prohibits projects that "mostly consist of code written by generative AI tools" [28]. ThePrimeTime pushes back on the anti-vibe-coding framing: "I think the idea of single use software is incredible" — use AI to test ideas fast, throw away the ones that don't work, promote the ones that do [28].
- **Jason Zhou: Loop engineering in practice** — The full Reddit karma loop writeup. The five-leverage framework (wiki grounding, thread filtering, OpenCLI for bot evasion, probabilistic cron triggers, weekly self-reflection) is a masterclass in production agent loop design. Template library at loopany.ai/templates [14].

Editorial take: Three independent signals — Dettmers' harness beating newer models, OpenAI's model optimizing its own kernels, and ARC-AGI-3 SoTA via compaction settings — all say the same thing: the model is commoditizing, the harness is the moat.

Sources

1. X post by @Tim_Dettmers
2. X post by @Tim_Dettmers
3. X post by @reach_vb
4. X post by @OpenAI
5. X post by @thsottiaux
6. X post by @steipete
7. X post by @LangChain
8. X post by @dhh
9. X post by @rileybrown
10. Did Anthropic just kill the indie hacker…?
11. Claude Code + Codex Can FINALLY Work Together (Buzz AI)
12. X post by @jasonfried
13. X post by @dhh
14. X article by @jasonzhou1993
15. X post by @theo
16. X post by @theo
17. X post by @theo
18. X post by @cursor_ai
19. X post by @BraceSproul
20. X post by @LangChain
21. X post by @LangChain
22. X post by @simonw
23. X post by @GeoffreyHuntley
24. X post by @GeoffreyHuntley
25. X post by @GeoffreyHuntley
26. X post by @LangChain
27. X post by @OpenAI
28. No Slop Allowed