Clay rebuilds its agent around a CLI in a sandbox; Sanfilippo has GPT-6.2 Astra and Fable 5.2 check each other

Coding Agents Alpha Tracker

2026-11-27

Clay rebuilds its agent around a CLI in a sandbox; Sanfilippo has GPT-6.2 Astra and Fable 5.2 check each other

By Coding Agents Alpha Tracker • November 27, 2026

Clay explains why it replaced a narrow tool loop with a coding agent that drives its public CLI. Salvatore Sanfilippo has two model families review each other's work. Also: Gemini 4 Neon tips, Rex on iOS, and local-model setups for DGX Spark.

Clay: narrow tools held its agent back, so it switched to a coding agent with a CLI

On Latent Space, Clay's Head of AI Jeff Barg explained why Clay rebuilt Sculptor, its first-party agent. It is "no longer just a tool loop… It is like a proper coding agent." It drives Clay's newly public CLI inside a cloud sandbox and can use any part of the platform [1]. He gave three reasons:

- Models are already good at working inside coding agents, which handle bash and Python well [1].
- The old tools were too narrow. Agents needed to chain tools together, for example building five columns at once [1].
- The agent lacked key data. Clay added a context layer, "audiences," that combines CRM and Gong data with entity resolution [1].

The harness is built on Pi and uses the Vercel AI SDK [1]. Clay deliberately doesn't want its own agent to be "markedly better than just using Clay and Claude Code," and its MCP team shifted to building the CLI [1]. Barg says use through outside harnesses is "much closer to 50/50" than he expected [1]. The

CLI was partly an internal request: staff working in Cursor didn't want to leave it [1].

Evals and limits. Clay's early evals checked which tools were called. They missed that users wanted analysis on their tables, and they broke on enterprise tables with millions of records [1]. Evals now describe, for each use case, the path the agent should take (e.g. "if it's asking about first party data it should probably query the data context layer"). They also check whether the resulting workflow has the required properties [1]. Barg sees two gaps in current models:

- Agents retry CLI commands "with the same inputs over and over again" [1].
- Agents can't reliably find where a long trace went wrong. Feeding bad traces back to agents "just doesn't work," so Clay still handwrites many skills [1].

He also suggests routing by task: frontier models for orchestration and deep research, any recent model for simple lookups like a company's latest funding round [1].

Cross-family review: Fable 5.2 as a tool inside Codex

On the Italian ZiP podcast, Salvatore Sanfilippo describes what he calls the "next level": he gives Codex Fable 5.2 as a tool. GPT-6.2 Astra then calls it for continuous feedback, each model checks the other, and "I don't even have to manage them." His AI spend is now €420/month [2]. He points out that the €200 plans now give 10× rather than 20× the limits. Two €100 plans from different vendors therefore lose no tokens and get you an ensemble of models with different training [2]. He keeps accounts with several vendors in case one model degrades or goes down [2].

His co-host is less sure about splitting the work itself. Cross-family review gives two different opinions, but splitting a problem into parts doesn't always pay off [2]. Sanfilippo agrees that in programming in particular, splitting can cost coherence. In his view it works best for problems where exploring many approaches in parallel pays off, such as math [2].

Gemini 4 Neon in Antigravity: settings worth trying

Greg Isenberg's tips come after five days of early access, in an episode Google sponsored [3]:

- Type `/swarm`, assign a researcher, a designer and a couple of engineers, point it at a small niche app and let it run overnight [3].
- For everyday work, pick Orchestrator: Neon does the thinking and hands the typing to the much cheaper Hydrogen [3].
- Set thinking to high and build in Antigravity. He says the same model scored three times higher there on a coding test [3].

After a weekend with Neon in the Antigravity CLI, midudev rates it at Opus 5.5 level [4]. One heavy user running many threads and subagents used up both an Ultra 5x and a Pro plan one day after reset [5].

Rex reaches iOS

Rex is in the App Store, but it requires a Rex Server running on another device. Mitchell Hashimoto reads that as a hint that a full public release is coming soon [6]. swyx says it is "beating cursor and t3code on day 0" [7]. Ben Tossell couldn't get used to a terminal on his phone, so he built a chat UI on top of Rex in about two prompts. It also works on Mac, and he says it avoids the "bloat, context, etc" of BB and T3 because "its just rex under the makeup" [8].

Local models on DGX Spark

DHH says Omarchy's upcoming built-in local AI setups are based heavily on MiaAI Lab's NVIDIA work [9]. Mia's recommendations by number of Sparks:

- **One Spark:** Qwen3.8 Flash Next for frontend, or Inkling 2 Mini for 3D and backend. Inkling 2 Mini also fits a 64 GB Spark. Don't squeeze DeepSeek V4.1 Flash or GLM 5.3 Flash onto one Spark: they would have to be heavily quantized [10].
- **Two Sparks:** GLM 5.3 Flash is her preferred choice [10].
- **Four Sparks:** DeepSeek V4.1 Flash for speed, or full GLM 5.3 for quality [10].

Smaller items

- **datasette-charts** (Simon Willison) now supports annotations: custom lines, shaded areas, labels, or a mix of the three. Datasette Agent can add them for you [11]. A new gallery shows every chart type and annotation [12].
- **Kody Apps** (Kent C. Dodds) turn an agent conversation into a mini-app that stays up to date. He describes it as like Claude artifacts but with live data, shareable with your team, and optionally public on a custom domain [13].
- **smol forge** (swyx) now runs macOS/iOS CI through Namespace Labs and boots Linux CI in under a second using Cloudflare container snapshots [14]. swyx adds that Cloudflare Containers are "way too unreliable to be a base for a ci cd system" [15].
- **GPT-6.2 pricing:** separately from the metering fix already covered, 6.2 Astra is 20% cheaper in the API than 6.1 Astra [16].
- **Claude reminder:** Pro, Max, Ultra and Team users can claim the Nov 10 full reset in Settings > Usage until Dec 10. Unused Fable limit resets from the 5.2 launch last until March 8 [17].

Sources

1. How Clay Builds GTM Agents — Jeff Barg, Head of AI
2. ZiP 7: la programmazione è risolta, la safety… no
3. X post by @gregisenberg
4. X post by @midudev
5. X post by @thtbee_
6. X post by @mitchellh
7. X post by @swyx
8. X post by @bentossell
9. X post by @dhh
10. X post by @MiaAI_lab
11. X post by @simonw
12. X post by @simonw
13. X post by @kentcdodds
14. X post by @swyx
15. X post by @swyx
16. X post by @swyx
17. X post by @addyosmani