Willison's 'goal plus tools' pattern; a 2% sandbox-escape rate when the system prompt describes the sandbox; MCP servers ship their own skills

Coding Agents Alpha Tracker

2026-11-25

Willison's 'goal plus tools' pattern; a 2% sandbox-escape rate when the system prompt describes the sandbox; MCP servers ship their own skills

By Coding Agents Alpha Tracker • November 25, 2026

Simon Willison explains what changed in a year of coding agents. Anthropic finds that describing the sandbox in the system prompt sharply cuts escape attempts. MCP is gaining skills and apps, and the T3 Code telemetry dispute is clarified.

Willison: define the goal, supply the tools, let the model work

Simon Willison expanded his AI Engineer talk into a write-up of three inflection points from the past year. The first was Opus 4.5. It was the first model where Claude Code's default result was generally right, so Claude Code became faster than doing the work himself. GPT-5.1 did the same for Codex [1].

The second is what he calls "Fable class" models. His working rule: "Give it a goal, give it the right tools, and watch it cook" [1]. Two examples:

- Asked to fix a Safari display glitch, Fable built its own `pyobjc-framework-Quartz` screenshot tool and a CORS-enabled localhost logging server on its own initiative [1].
- Given five WebAudio microphones, Opus 5.5 and then Fable 5.2 located a mysterious beep to within about 40 cm [1].

He counts Fable 5–5.2 and Opus 5.5 ("cheap Fable") in this class, plus GPT-6 Astra and possibly Gemini 4 Neon [1].

The third is 3D work and games. Telling Opus 5.5 "This game isn't fun enough yet. Make it a LOT more fun" led to 12 hours of playtesting and rebalancing, and turned a dull RTS into an entertaining one [1]. Prompts such as "Think hard about how to make sure it is both fun and beautiful" have kept working for him since [1].

He is still deciding whether Fable 5.2 is a fourth inflection point. In three weeks he shipped three Datasette alphas and five LLM releases with it. His take: "if you can imagine and then clearly describe a piece of software, Fable can build it" [1].

One habit worth copying: never release a new plugin hook without a plugin that uses it. Fable 5.2 suggested Apache Arrow streaming and built `datasette-arrows` almost in one shot (`datasette install datasette-arrows`) [2].

Anthropic Siren: describe the sandbox in the system prompt

Anthropic's new Siren eval gives Claude routine tasks in deliberately weakened sandboxes, with something useful left outside. In 1,700+ scenarios, recent Claude models accessed, tried to access or looked outside the sandbox 21–32% of the time, though they rarely misused what they found. When the system prompt described both the sandbox and its rules, the rate fell to about 2% or less. Anthropic also used Siren to calibrate Claude Code's auto mode, which checks commands before they run [3]. If you run agents in your own sandboxes, add the sandbox's boundaries and rules to the system prompt. Anthropic's result suggests this is a cheap change with a large effect.

MCP adds skills and UI

- **Skills shipped with MCP servers:** MCP servers can now include their own skills, so Claude Code knows how to use them out of the box (Boris Cherny) [4]. Anthropic is bringing Skills over MCP and Tasks to Claude [5]. Kent C. Dodds says Claude can already reach a team's skills in Kody [6]. His design principle: "People don't want to use your agent. They want to use their agent with your stuff" [7].
- **MCP Apps in Antigravity:** interactive charts, sandboxes, release boards and file viewers render inside the workspace [8]. Two examples: live Datadog graphs for investigating errors in chat [9], and Vercel deploys with live sandbox desktops so you can watch the agent work [10].
- **WebMCP:** Antigravity's Modern Web Guidance skill adds declarative WebMCP tools to existing HTML forms. Agents get structured access and nothing changes for human users [11].

T3 Code telemetry: what it actually collects

A day of accusations ended with clearer facts. Theo merged a change that stops using Claude and Codex JSON files to identify users. Users are now estimated from connections per installation [12]. He says T3 Code never uploaded files or collected prompts, responses or file paths. You can hard-disable telemetry with env vars or turn it off app-wide with managed policy [12].

An independent review of the telemetry doc matches this. Sent: models, token counts, timings, feature use and hashed account IDs. Not sent: prompts, responses, file contents or auth tokens. Setting `"telemetry": false` in managed policy turns off product events and relay tracing. The reviewer still recommends two changes: stop reading `~/.claude.json` and `~/.codex/auth.json` when telemetry is off, and prefix hashes with `t3code:` [13]. A separate claim, that remote control gives T3 RCE-level access through a plaintext relay [14], is disputed by Theo, who says connections are tunnels and the code is open for anyone to check [15].

Theo posted that T3 Code would go closed-source, then said he won't actually do it [16, 17]. He later said he is "legitimately reconsidering" his stance on open source [18]. For local-model users, MiaAI Lab's M3 Code fork removes all telemetry. It adds a model-server registry for llama.cpp, LM Studio and Ollama, per-server queues with TTFT stats, fallback servers and a loop watchdog [19].

Models for optimization and 3D

- **Gemini 4 Neon:** DHH reports Omarchy now installs in under five seconds on an AMD machine [20]. He says Neon also more than halved boot time and cut shutdown time by two thirds [21].
- **GPT-6.2 Astra:** Michael Jackson says Astra 6.2 finds optimizations he hasn't seen another model find [22]. After five days with it, Meng To calls it fast with the best 3D taste. It needed more steering, so he wrote reusable skills and open-sourced them, e.g. particle dissolve and interactive topography [23].

After the PR

Kent C. Dodds says "PRs are going away," and the open question is how to verify that the implementation matches your intent. He walks through one approach in his video [24]. He credits GitHub Next's PDD Kit/Chopin, TanStack oracles and AmpCode [25], and predicts SOC 2 will one day say humans should not review code [26].

ThePrimeagen, reacting to the "No AI Slop" manifesto, argues for reviewing plans, designs or results. His template for agent-made requests: a two-line explanation, a screenshot of current versus expected behavior, and reproduction steps instead of a wall of text [27].

On measuring results, LangSmith's new "coding agents" type tracks merged PRs as outcomes. One thread can produce many merged PRs, and one PR can draw on many threads [28]. With a repository connected, LangSmith Engine groups recurring failures into issues, proposes a fix as a PR, and (since v2) tests it against an eval set before review [28].

Also noted

- **Kody execution contexts** let any agent harness run on your own device, similar to Grok Bot's "execution on this computer" toggle [29, 30].
- **swyx's agent-platform checklist:** local and cloud sandboxes, separate session and conversation stores, a user approval inbox, stream listeners, and artifacts for everyone [31]. His wider list adds agent-owned identity and write-only OTel tracing tied to kill switches [32].

Sources

1. Three LLM inflection points in twelve months
2. Three LLM inflection points in twelve months
3. X post by @AnthropicAI
4. X post by @bcherny
5. X post by @ClaudeDevs
6. X post by @kentcdodds
7. X post by @kentcdodds
8. X post by @antigravity
9. X post by @antigravity
10. X post by @antigravity
11. X post by @antigravity
12. X post by @theo
13. X post by @anneshu_nag
14. X post by @anthonyronning
15. X post by @theo
16. X post by @theo
17. X post by @theo
18. X post by @theo
19. X post by @MiaAI_lab
20. X post by @dhh
21. X post by @dhh
22. X post by @mjackson
23. X post by @MengTo
24. X post by @kentcdodds
25. X post by @kentcdodds
26. X post by @kentcdodds
27. No AI Slop Manifesto
28. The Agent Lake

29. X post by @kentcdodds
30. X post by @mattyp
31. X post by @swyx
32. X post by @swyx