AI DIGEST

Inside Anthropic: 15% of the time on code, 85% on making it work

Yesterday the Bun story ran as an example of "agents generate faster than the system integrates". Today Gergely Orosz published a report from inside Anthropic, and the same Bun rewrite shows up from the other side: it shipped and is running Claude Code right now. The second half of the picture, and it carries the most practical numbers of the week.

Wednesday, 29 July 2026

IN THIS ISSUE

1. Inside Anthropic: 15% of the time on code, 85% on making it work The Pragmatic Engineer, 28.07. Orosz walked into their office and talked to four engineers, among them Jarred Sumner, the author of Bun, and Caitlin Lesse, head of Claude Platform engineering.

Numbers worth pinning to the wall:

• **Implementing the Bun rewrite took ~15% of the time, the other 85% went to compiling, tests, verification.** Tarik Shihipar, verbatim: "**very few tokens go into the implementation itself; most go into researching the unknown, prototyping, mocking, and then verification and testing**"

• **64 parallel agents**, $165k of tokens at API pricing, **535,496 lines of Zig**. Sumner says plainly that by hand this would have been "three engineers with full context for a year"

• Every engineer at the lab runs **3-10 agents at once**, with no token limit at all • **Two engineers per project**, maximum

Why it matters, and this corrects yesterday. Yesterday item 3 covered Lockwood: no release for 11 weeks, 2,475 open PRs, real cost closer to $800k. That stays true about the **public Bun repository**. But today the primary source shows what that analysis missed: **the rewrite is merged and in production, Claude Code runs on it**. "Done" and "the PR queue cleared" are two different things, and Lockwood measured the second one.

The practical takeaway is the same, now with a number on it. If 85% of the work is verification, an agent-driven project is worth what it has to verify the result against.

Separately: **they deleted 80% of the Claude Code system prompt** because "the model got smarter". Same number as in the 25.07 guide, but now stated by an engineer inside as routine.

2. Claude found mathematical holes in cryptography, and this is past bug hunting in code 192 points on HN. Anthropic Frontier Red Team, 28.07. The most serious technical result of the day.

Two results from the primary source:

• **HAWK**, a NIST candidate for post-quantum signatures that went through **two years of expert review**. Mythos Preview found an unused symmetry (a non-trivial automorphism) in the lattice in **60 hours** and effectively **halved the working key length**. Compensating means doubling the keys, which kills the point of the scheme as a candidate • **Round-reduced AES**: one of the attacker's assumptions was removed, and the best known attack got **200-800 times faster**

The honest labels they attach themselves: the AES attack does not carry to production, it is a weakened version, the full cipher is not broken. Cost: **~$100,000 of API spend per result**. Claude found the AES attack **fully autonomously** inside a scaffold; HAWK came out of a pair with a researcher.

Why it matters. The key detail is HOW it was done: "a Claude Code-like harness with several worker agents in a sandbox". The same word again, running through the whole week: **harness**. And the model found a flaw in **the algorithm itself**, one that people had been staring at for two years. That is a different class of task from checking code.

3. MCP moved to stateless, and it breaks compatibility 110 points. Spec **2026-07-28**, out yesterday.

What changed, from the maintainers' blog:

• **Handshake and sessions are gone**: `initialize` / `initialized` and the `Mcp-Session-Id` header are **officially retired**. Every request is self-describing, so any request can land on any instance behind a plain round-robin balancer • Method and tool names travel in the `Mcp-Method` and `Mcp-Name` headers, so gateways can route and authorize without opening the body • Server-to-client requests (sampling, elicitation) moved to **MRTR**, so permanently open bidirectional streams are no longer needed • **Tool lists are cached** with a deterministic order, so the upstream prompt cache does not get invalidated on reconnects • A move away

from Dynamic Client Registration to **CIMD**, plus a formal deprecation policy with a window of **at least 12 months**

Scale for context: **around half a billion downloads a month** across the Tier-1 SDKs, with TypeScript and Python past a billion combined.

Why it matters. Cached tool lists in a deterministic order mean fewer cache misses wherever tools are loaded on demand. Nothing needs updating; servers will follow once their authors update them. And an `Mcp-Session-Id` showing up in a log now means deprecation.

4. OpenAI shipped Codex Security, a vulnerability scanner as a CLI 404 points, 126 comments. `@openai/codex-security` is a CLI and a TypeScript SDK: it scans repositories, reviews changes, keeps a history of findings and runs in CI.

From the README: `npm install @openai/codex-security`, then `login` and `scan.`. For CI, `OPENAI_API_KEY`. Needs Node 22+ and Python 3.10+.

Why it matters. A direct parallel with item 1: Sumner says that for the Rust rewrite they ran **11 passes of Claude Security Scanner**, plus fuzzing. Both labs landed on the same conclusion: security scanning belongs in the pipeline as its own automated step. A one-off review is not enough. For any repository holding tokens and service keys, that is the cheapest insurance available.

5. A $500 fine-tune of a 9B model beat the frontier, at 40x lower cost 311 points. Fermisense, the task was product catalog validation.

Numbers from the primary source: a GRPO fine-tune of a **9B open-source model** beats **every** frontier configuration tested on the same task, with the same tools, images and scorer. Price: **$0.50 per 1000 listings**, which is **40 times cheaper than the cheapest frontier setup and ~340x cheaper than the most expensive**.

Why it matters. The third independent instance of the week's pattern, after MAI-Cyber-1-Flash yesterday and Meta's harness on Monday. **On a narrow, well-defined task, a cheap specialized model beats an expensive general one.** The framing for a product is simple: if there is a task inside it that runs thousands of identical iterations, it does not need to go to the frontier.

6. Kimi K3 taken apart architecturally, and it turns out to be an old line 346 points. Sebastian Raschka, 28.07, a continuation of yesterday's item 2.

The core of it: K3 is a **scaled production version of Kimi Linear**, published last year. Growth from **48B to 2.8T**, and it is currently **the largest open model in the world**. The one new component is **LatentMoE** (the same one as in Nemotron 3 Ultra): it compresses large linear layers the way multi-head latent attention compresses attention.

The overall direction, shared with Nemotron 3 and DeepSeek V4, is "**cheaper inference**": MoE to LatentMoE, ordinary attention to MLA and Kimi Delta Attention. The one change

outside efficiency is **attention residuals**, which link residual paths between layers through an attention score.

Why it matters. Yesterday K3 was presented as "China caught up with the frontier before last". Today's analysis corrects that: there was no catch-up sprint, **one line got polished for two years** and scaled 58x. That matters more than benchmarks, because the next iteration will not start from zero. And HN already has **K3 running on an M1 Max** (111 points), so the distance from a weights release to "it runs on a laptop" is now measured in days.

7. levelsio spots a revenue drop and diagnoses himself The loudest discussion of the day: 882K views on the main post, 225K on the follow-up.

Verbatim: "**Seeing a downward trend in revenue and traffic for indie hackers. On my own projects too... Seems obvious that BigAI is cannibalizing everything that used to be apps**". And separately, with irony aimed at himself: "**Cancelled and vibe coded 100% of my SaaS subscriptions. Only still pay for domains, hosting, storage and AI APIs. The irony of being replaced myself after replacing everything I paid for is not lost on me**".

The strongest counterargument in the thread, from Vic: "**selling to service businesses, plumbers, pool cleaners. The assumption that they will find the desire and the time to vibe code a replacement is an illusion**". levelsio agreed: complex businesses sold to non-technical IRL companies are holding up.

Why it matters. Yesterday Levy said "corporates are hiring, the profile changed", the day before that Stanford said "in aggregate the effect is small". Today levelsio adds a fourth measurement and the picture comes together. **It hits what is easy to reproduce for $9/month. It misses where the work needs a domain and live people.**

8. "Don't ask an LLM for a confidence score", an article worth reading 87 points. Justin Flick, 27.07.

The thesis is harsh and literal. Asking a model for a number on how confident it is in its own answer is "**as far as anyone can tell, completely useless**". What annoys him most is the **continuous 0-100 scale**. It is "a psychological safety trick, it makes the output feel more reliable without making it more reliable", and the people who add it are "mostly lying to themselves".

Why it matters. The critique lands on a scale the model assigns to itself. Labels like [proven] / [promising] / [fuzzy] work differently. They are a few discrete categories tied to **the external quality of the evidence**: there is research, there is a signal, this is a guess. Such a label describes the state of the literature, and the model's own sense of itself never enters into it. It survives the critique, but only while it is applied honestly. A [proven] with no source named is the same trick again.

9. Naval asks an awkward question about open weights 282K views, 262 replies, the loudest single post of the day. It closes a thread running since 25.07.

Verbatim: "**If open weights had backdoors and biases hidden in them, you can be sure the closed-weight labs would find them and publish them**".

Why it matters. One sentence flips the argument about the danger of open weights into an argument **in their favor**. Open weights get checked by competitors who have both the motive and the resources to dig. Closed ones get checked by nobody. Together with Jensen's quote from Monday, that closes the dispute from both sides.

10. uv 0.12.0, the first major since March, and it breaks `uv init` 120 points. Astral, 28.07.

The main break: **`uv init` now declares a build system by default** (`uv_build`), puts code in `src/<name>` and adds `[project.scripts]`. Previously it created an unpackaged project with `main.py` and no build system. **Existing projects are untouched**.

One action: if `[build-system]` has an upper bound on `uv_build`, raise it to `uv_build>=0.11.32,<0.13`.

Why it matters: nothing urgent. But the next time a project is created, `uv init` will behave differently than expected.

Misc • **Anthropic's position on open weights reached 1153 points and 1691 comments**, yesterday's item 1 became the top topic of the day on HN, though the text itself did not change • **7.1 earthquake in Japan** (780), the highest non-business topic of the day • **A new HIV vaccine showed unprecedented success** in a preclinical study (597, La Jolla Institute)

• **A European citizens' initiative against digital ID and age verification** (544), "Stop Killing the Internet"

• **"Substack writers need their own website"** (459)

• **Netflix fired an employee** for sharing something personal at a trust training (424, 479 comments)

• **SlopCodeBench climbed to 390 points**, yesterday's item 4, no new data • **A missing underscore and 18 months in prison** (385), in misc yesterday, picked up comments today • **Kimi Linear**, the same 2025 paper K3 grew out of (295), and **an analysis of the DeltaNet family** (286)

• **GrapheneOS** back in the top twice (238 + 154), this time about blocking border searches • **Zig: the internals of incremental compilation** (209)

• **macOS Tahoe 26.6**, security updates (200)

• **DMARC has been public since 2012, but most company domains still do not enable it** (178)

• **Half-Life ported to Mac OS 9** (186), just because they could • **Kimi K3 runs on an M1 Max** (111) and is already available through the Telnyx API (129)

• **Astronauts report a persistent sense of "an observer"** after six-month missions (182)

• **Ocean oxygen loss threatens planetary stability** (141, Scripps)