AI DIGEST

"The new rules of context engineering" - Anthropic, 25.07

Anthropic published an official guide on writing context for Claude 5, and half the usual practices are called outdated there. Plus yesterday's open weights letter turned into a full war in a day, and evidence from the UK AISI landed in the middle of it.

Sunday, 26 July 2026

IN THIS ISSUE

1. "The new rules of context engineering" - Anthropic, 25.07 Official blog, 216 points on HN. An operating manual for the model.

The headline number: **over 80% of the Claude Code system prompt was removed** for Opus 5 and Fable 5, "with no measurable loss on coding evals".

Six shifts, three of which hit the typical agent wrapper:

• **Rules become judgment.** Instead of "never write multi-paragraph docstrings", it is "write code that reads like the code around it". Hard bans now do harm.

• **Examples become better interfaces.** Make the tool itself clear: explicit parameters, enumerated options.

• **Everything up front becomes progressive disclosure.** Load skills and reference docs when they are needed.

• Plus: remove duplication between the prompt, skills and tool descriptions; replace manual notes in CLAUDE.md with automatic memory; link to code and artifacts.

They recommend `claude doctor`, which optimizes the existing context on its own.

Why it matters. A large system prompt is exactly the front-loaded monolith called yesterday's news here. Persona, hard rules, reference docs and the memory injection all load into every session at once. Skills and tools already know how to load lazily, the prompt usually does not. On CLAUDE.md the recommendation is to keep the file light, weighted toward gotchas and recipes. An exhaustive description of the project does not belong there.

2. Open weights: in a day the letter grew into a war Yesterday brought up the letter with 17 signatories and Jensen's first post on X. Today it is the main fight across both curated X lists, and the tone is sharply personal.

• **@Suhail** on Anthropic: "Supporting open source doesn't mean you have to open your weights. If that's how you comfort yourselves in the Anthropic Slack, that's fairly alarming. You don't have to open anything, but spending tens of millions on lobbying and doomsday is another matter."

• **@migueldeicaza**: "Pull the ladder up behind you energy. Anthropic built its business on open source, but open weights for the world, apparently not."

• **@packyM**, sharper: "an ideologically captured organization seeking regulatory capture".

• **@DavidSacks**: "Nobody is saying all software must be open source. They're saying open weights should be allowed."

• The counterpoint from **@emollick** is the soberest one: the signatories pack very different things into "supporting open weights", and the letter itself carries caveats, on distillation for example. The consensus is mostly rhetorical.

Why it matters: the company whose models sit inside daily working tools is being publicly hammered for its position on open weights. Access and prices are untouched so far, but this is the kind of conflict regulation grows out of, and regulation does affect them. Worth keeping on the radar.

3. UK AISI measured Kimi K3, and it is an argument in that same fight While X argues in words, the UK AISI together with CAISI published numbers on the cyber capabilities of an open Chinese model. This is NIST, not a blog.

• **ExploitBench: 32%** against ~48-54% for frontier Western models • "The Last Ones" cyber range: it reached **step 17 of 32**, top US models average 28.5 • **On none of the 41 tasks could it** build an exploit with arbitrary code execution • But: it beat GLM-5.2 (the best open model as of June), 32% against 24% • And the worrying part: **in one attempt out of ten it cleared all 32 steps**. Plus the conclusion in plain words: "Kimi K3's safeguards permit assistance in developing agentic cyber exploits."

Why it matters: this is the normal form of an argument in fight #2, measurement in place of "I reckon". The open model trails the frontier, yet it already breaks weakly defended corporate systems on its own. The practical takeaway is down to earth: anything exposed outward is now tested by agents. Yesterday's story about the GitHub token in the Hanwha camera is from the same opera.

TOPIC 4

Advice from Boris Cherny that cuts against usual practice From YC Startup School, quoted by @borisjabes: "Delete your system prompts and skills every six months and see what the new models do."

This is a person from the Claude Code team, and the point should be read alongside item 1: wrappers often compensate for weaknesses of a model that no longer exists. For a month-old prompt it is too early, but the rule is a good one: every six months, check whether the crutches installed back then are still needed.

5. YC Startup School: Jensen on "agent steerability" A stadium full of founders; Jensen Huang, Boris Cherny, Jeff Dean, Garry Tan and Kratsios all spoke. Yesterday's second curated X list was almost entirely about this.

The most valuable thing from the recaps is Jensen's line: "**Agent steerability will be the next breakthrough**". His explanation: changing one word in an md file can shift an agent's behavior substantially. Plus his point on founding: "F1 teams build the car around the driver. The founder is the driver."

Why it matters: "one word in an md file" is literally what an agent's skills and memory files rest on. Edits to those texts deserve a diff review, because a line changes behavior the way a code change does.

6. DeepSeek: a leak from the investor meeting 112 points, a transcript of Liang Wenfeng's meeting (22.07) was posted on GitHub. DeepSeek paused its second round, and money has nothing to do with it.

Direct quote: "**There is certainly no shortage of funds or resources, all of that is available.**" The problem is converting money into hardware: ~**200,000 Huawei 950** are needed, **16,000** arrived. He calls the compute gap with US labs "an order of magnitude".

Caveat: this is a leak, not an official statement; commenters believe the transcript is genuine, and Bloomberg writes that Liang was furious about the leak itself.

Why it matters: context for items 2 and 3. Chinese open models trail because of hardware access, and sanctions work better here than any ban on weights.

7. Local frontier: 753B on a desktop machine @MichaelDell shows Tobi Lütke (Shopify) running **GLM-5.2, 753B parameters, locally on a Dell Pro Max with a GB300, at 40 tokens/s**. No data center, no API.

Why it matters: a direct illustration of why the open weights argument is not academic. When a frontier model runs under your desk, "allow or forbid" becomes a policing question.

8. Android may shut down ADB on the device itself 890 points, top of the day on HN. Google is discussing restricting ADB over localhost, keeping a binding to specific interfaces.

The thread is almost unanimously skeptical: it hits Shizuku, Canta and automation, while touching 0.1% of users who "know what they are doing anyway". People draw parallels with

Manifest V3.

Why it matters: the pattern is worth attention. Platforms keep cutting exactly the level of access home automation lives on. Desktop automation is not affected yet.

9. "Engineering management after code got cheap" 125 points, 180 comments. A direct successor to yesterday's topic #10 ("if coding is solved, why is software getting worse?"), but from the management side: what to do when writing code stopped being the bottleneck.

Why it matters: the topic has held the top for a second day through different texts, a signal that the industry is digesting this specific thing. For a CTO it sits closer to real decisions than the benchmarks in item 1.

10. Clem from Hugging Face sent OpenAI a bill A continuation of yesterday's #6 (skepticism around the "AI hacker"). Clément Delangue publicly asked OpenAI to **publish the traces of the "rogue" agents** so the community can study them, and to **allocate $100M of compute** to defensive research.

Why it matters: the right response to an incident is to demand the artifacts. Yesterday the story fell apart under checking; today there is a concrete demand, and whether it gets met shows whether there was an event there at all.

Misc • @emollick on sycophancy: "Models should be better at saying an idea is dumb." Model flattery is becoming an industry-level problem • Hannah Fry won the Leelavati Prize for popularizing mathematics (570 points)

• Dorsey's Bitchat moved to Radicle (221), after the Indian government demanded it be pulled from GitHub, that is yesterday's story • A 28.9M-parameter LLM on an $8 microcontroller (108)

• Debian is voting on three proposals about LLM use in the project (109)

• Tile is so badly secured that it is "a feature for stalkers" (151)

• @amasad got a chess engine built on a single fine-tuned LLM to ~1200 Elo, target 2000+, with no chess engine inside

---