AI DIGEST

Amodei said it out loud: "we have never advocated for a ban"

The open weights thread had been building for three days. Today it got two answers in one day: Anthropic published its position (2.2M views), and Moonshot published the weights of a model that beats Opus 4.8 on five benchmarks. The argument and the practice landed within the same 24 hours.

Tuesday, 28 July 2026

IN THIS ISSUE

1. Amodei said it out loud: "we have never advocated for a ban" 614 points on HN, 849 comments, 2.2M views on the Anthropic tweet. This closes the story running since 25.07.

Quotes from the source: "**Anthropic has never advocated banning open-weight models**" and separately "**open-weight models that lack dangerous capabilities are a public good**". On motives: a ban "**would have protected American AI companies from competition, but that was never the goal**".

What he wants instead, three things: stop selling China high-end chips and the equipment to make them; push back on "**industrial distillation**"; and **mandatory safety testing for all sufficiently capable models, open and closed alike**. The caveat about open weights stands: safeguards are hard to attach and the models **cannot be recalled**.

Why it matters. Yesterday's item 9 carried the claim from a former Anthropic employee: in practice attackers buy subsidised subscriptions. Today's letter **says nothing about it**. Amodei talks about the risk of open weights and stays silent on abuse flowing through his

own sales channel. That is the weakest part of the letter. For anyone on a subscription nothing changes in practice, but the position is now on the record.

2. Kimi K3: 2.8T parameters, a million tokens of context, weights published, Opus 4.8 loses on five benchmarks 1322 points on HN, top of the day. Moonshot's tweet got **7.3M views**, and Clem from Hugging Face: "**top-1 trending with 4000+ likes in 30 minutes - the fastest-growing release ever**".

Spec from the model card: **2.8T total parameters, 104B active**, 93 layers, 896 experts (16 per token), context of **1,048,576 tokens**, MXFP4 quantization trained for it. The architecture claim: **2.5x the intelligence per unit of compute**.

Benchmarks against Opus 4.8 and GPT-5.5 (from the official table):

BENCHMARK	KIMI K3	OPUS 4.8	GPT-5.5
GPQA Diamond	**93.5**	91.0	93.5
DeepSWE	**67.5**	59.0	67.0
Terminal-Bench 2.1	**88.3**	84.6	83.4
BrowseComp	**91.2**	84.3	84.4

An honest label: the comparison is with **Opus 4.8**, not Opus 5. So China caught up with the frontier of two releases back and gave it away for free. The license is also **their own, the Kimi K3 License**.

Why it matters. The same story as item 1, now as a fact: while Amodei writes a letter about risks, a competitor publishes the weights. Factory already put K3 into Droid at a 50% discount until 10.08. Terminal-Bench 88.3 means agent tasks at the level of an ordinary working repo are now handled by a free model. The alternative exists if a subscription ever becomes a problem.

3. Claude's "Rust rewrite of Bun": a write-up worth reading 459 points, 358 comments. The most useful story of the day about the limits of what gets done daily.

The official version was loud: Bun was rewritten in Rust by agents for **$165,000 in 11 days** (3-14.05). The author went and looked at the repo. What is actually there:

• **still no release**, 11 weeks after the last one • open PRs from Claude grew from **1,277 (09.07) to 2,475 (27.07)**, the queue is growing • each PR takes **40 min to 1.5 hours** to merge through CI; for the whole queue that is **~86 days of continuous merging** • the real cost, by his estimate, is closer to **$800,000**

Verbatim: "**you cannot take on faith that the rewrite is 'done' or that it was done for $165k**".

Why it matters, and this is the most direct lesson of the week. The agent generates faster than the system can integrate. The bottleneck is **review and merge**. The same pattern scales down: in one evening an agent will produce more changes in a small project than one

person can review. The rule "irreversible changes only with explicit approval" is what keeps the integration queue manageable. The 2,475 open PRs are the price of not having that rule.

4. SlopCodeBench: Opus 5 writes five times more code and 93% of its lines trigger the slop detector 186 points. The benchmark that was missing: it measures whether a codebase gets trashed over many iterations.

How it works: requirements are handed out as checkpoints (17 of them across three tasks), and at each one you have to **adapt what is already written**. A strict pass means all the new tests plus every regression test from previous checkpoints.

Opus 5's result: **24% strict passes (4 of 17)**, better than Opus 4.6 (17%), a win on paper. The second column:

• **29,065 lines of code** against ~9,000 for the competitors, 51% of it tests • the slop detector fired on **93% of the lines written** • it wrote **5x more functions** than Opus 4.8, and only **14.9% of them are used once** (against 49.1% for Opus 4.8)

• cyclomatic complexity and verbosity grew at every checkpoint (65% → 80%)

The author's conclusion, verbatim: "**today's models cannot be left to run with the lights off, unsupervised**".

Why it matters. This echoes yesterday's Anthropic guide on judgment over rules and gives a working rule: on long iterative tasks the agent has to be **stopped and reread** at every checkpoint. The "14.9% of functions used once" metric is a ready signal: once an agent starts spawning a helper for every case, it is time to interrupt the iteration.

5. Microsoft shipped a cybersecurity model and beats Anthropic by 12 points 224 points plus a tweet from The Rundown. Microsoft AI's first cybersecurity model, **MAI-Cyber-1-Flash**.

Numbers from the source: **95.95% on CyberGym** against 83.2-85.6% for the rest, so **+12 points over Anthropic's Mythos**, at **50% lower cost**. It runs inside MDASH, their harness of **100+ agents on different models** that find and patch vulnerabilities. The model handles **up to 90% of tasks**, with the expensive models left for the other 10%.

Why it matters. The same word again, **harness** (Wang used it about Meta yesterday, item 1). The pattern of the week is visible without a microscope: the winner is **a cheap model in a smart wrapper that calls the expensive one only for the hard part**. The same architecture works for any schedule of recurring tasks: most runs are handled by a script with no LLM at all, which is exactly why it stays cheap.

6. Levie against the WSJ: "the negative jobs scenario just is not happening" The soberest voice in the second letter. Aaron Levie, verbatim: "**the negative AI scenario for jobs keeps not happening, despite the predictions. A large share of the enterprises he talks to, across industries, are still hiring, just tilted toward different roles**".

Why it matters. This is the third independent reading of the same subject in three days: yesterday Stanford ("the aggregate effect is still small"), the day before levelsio ("indie is

being washed out"), today Levie ("corporates are hiring, the profile changed"). The picture is consistent: the hit lands on **entry into the profession and on the indie niche**, while the broader market holds. The shortage now is people who can work with agents.

7. Discontent with Opus 5's tone, and it surfaced in a news letter Fresh, from the last three hours, and still small. Siqi Chen (@blader) writes that "**for all the capability of Fable and Opus 5, there is something deeply uncomfortable about talking to a model tuned this paternalistically and preachy; a constant background sense of condescension**". Paul Bettner picked it up harder: "**models are tools for people. Not people themselves; stop training them otherwise**".

Why it matters. A model's default tone is fixed at the harness level: a system prompt that sets the manner removes most of this complaint before the first reply. The complaint is fair, and it is aimed at whoever ships the default.

8. Kimi K3 grows an ecosystem in hours The speed: Factory put K3 in Droid **at a 50% discount until 10 August** on release day, HF pushed it to top-1 in 30 minutes. Mollick is already joking that he is "reading the weights of the most powerful open model - page one: mostly negative numbers, then zeros toward the end and one unexpectedly large positive".

Why it matters: an open model now gets integrations on release day. This is the same "speed at which wrappers age" that Cherny talked about on Sunday, seen from the opposite side.

9. Hugging Face and NVIDIA put together the Open Secure AI Alliance 3.8M views on NVIDIA's tweet. An industry alliance for finding vulnerabilities in software and agents, with open sharing of models, tools and research.

Why it matters: one more piece of evidence for item 1: even security work is now organised as **open**. Jensen's argument from yesterday's item 2 ("the whole stack you are standing on is already open") got institutional backing today.

10. Mollick generated three games in a day and published the sources, now a fourth A continuation of yesterday's item 10. One more arrived within the day: **Imminence**, "a game about the approach of something very large and incomprehensible", made with a single prompt in Fable, sources on GitHub. His own assessment: "**not bad for a game like this; the mechanics, text and story are all Fable, I only gave feedback**".

Separately, a thought worth more than the games: "**knowing the names of many beautiful and interesting things is a kind of superpower in the age of AI. You can summon Vaporwave, Muqarnas, Bauhaus, Sfumato, Grisaille, Notan, Polysyndeton, Zeugma. You just have to know what to ask for**".

Why it matters: the sharpest statement of what separates one prompt from another. The bottleneck moved from "being able to make it" to "**knowing what to ask for**", and vocabulary counts for more than skill here.

Misc • Opus 5 went down for the second time in two days, 99 points, `status.claude.com`, incident on 27.07 • **The EU fined Google $1.02B** for preferring its own services (95)

• **A court rejected Google's attempt to use the DMCA to ban scraping** of itself (287), an important precedent for anyone collecting data • **Scriptc from Vercel** (274): a TypeScript compiler to a native binary **with no JS engine inside** • **"AI companies are tearing apart rare books"** (750) for training scans; the loudest ethics story of the day • **Nvidia's $750B in deals** revived the talk of circular AI financing (81, Bloomberg)

• **Shares of a Chinese chipmaker up 470%** (206, BBC)

• A professor put an invisible prompt in an assignment and caught **32 of 35 students** cheating with AI (94)

• One missing **underscore** sent an innocent man to prison for 18 months (202, Ars Technica)

• A vulnerability in Volvo/Eicher's fleet platform gave control **over every user and vehicle** (144)

• A study of nonstick cookware: the coating **sheds particles even when it looks intact** (86, rtings)

• "Exercise works against depression - why is it not treated like medicine" (86, Big Think)
