AI DIGEST

Gemini 3.8 Flash and 3.8 Flash Cyber: the third Flash in six weeks, the price unchanged - and Google deliberately withheld exploit ability

AI digest, Thursday 03.09 - the third lab in three days drew a line on cyber capability, and drew it in a third place. While they measure each other on benchmarks, a small outfit found six real CVEs in curl where Mythos and Codex Security returned zero.

Thursday, 3 September 2026

IN THIS ISSUE

TOPIC 1

Gemini 3.8 Flash and 3.8 Flash Cyber: the third Flash in six weeks, the price unchanged - and Google deliberately withheld exploit ability

confirmed by: Google announcement · @GoogleDeepMind (**331.4k views** - the loudest post of the day across both lists) · HN 890 points, 519 comments · Ars Technica · WSJ "New Google AI Model Said to Narrow Gap on Coding Ability" (headline from RSS)

The key number is the one that did not move. Verbatim from the announcement: 3.8 Flash ships "at the same introductory price as 3.7 Flash" - **$0.75 per million input and $3.75 per million output tokens**. The third Flash in six weeks (3.7 Flash landed three weeks ago), and each time a gain with no price rise.

Claimed capability: on **DeepSWE v1.1** (long-horizon engineering) 3.8 Flash beats most larger frontier models "at a fraction of the cost"; **54.9% on HLE-Verified**; ahead of 3.7 Flash and "other frontier models" on Vals Finance Agent V2 and Harvey's Legal Agent Benchmark.

And right there, a caveat worth reading more closely than the benchmarks. Verbatim: the gain comes from 3.8 Flash **working harder** - taking extra reasoning steps, calling tools iteratively, and "at times the model may use **more tokens** to maximise performance, especially at higher effort levels". Stated plainly: if compute efficiency is the binding constraint, use a lower effort level **or stay on 3.7 Flash**, which remains fully supported. "Same price per token" is not the same as "same price per task".

Now the second model, and it is the more interesting one. 3.8 Flash Cyber goes only to "trusted defenders" through a new **Fairwind programme** (government cyber authorities, critical infrastructure operators, software maintainers). Numbers from their own text:

- **CyberGym** (the industry standard for vulnerability discovery) - frontier level, beating "significantly larger" models.
- An internal benchmark across **20 programming languages** (CyberGym is almost entirely C/C++) - success rate **over 70%**.
- **CWE-Bench** for patching (external, from Collinear) - **pass@1 47.2%** against 47.8% for the leader, but "at significantly lower cost".
- The **Chrome** security team: 3.8 Flash Cyber produced **2.6x more correct patches** than the best commercial models, which are "significantly larger".
- **Wiz**: +7.5-9.7% recall on their internal pentest benchmark at **2.3-5.2x lower cost**.
- The Cloud Vulnerability Research team found a **critical fundamental vulnerability in under 2 hours**, where this normally takes months.

And here is the sentence this item leads on. Verbatim: they "invested in vulnerability **remediation** from the start and **prioritised it over offensive capabilities such as exploitation**".

Following yesterday, and it now has three dimensions. In three days, three labs drew the line in three different places on the same phenomenon:

- **Anthropic** (01.09): the model may be used to **find** vulnerabilities, but **not to write exploits**.
- **OpenAI** (01.09): Astra **writes exploits end to end**, so it got a "critical" rating and access was narrowed to a small group.
- **Google** (02.09): offensive capability **deliberately not developed**, and the strong cyber version is not released publicly, only to trusted defenders through Fairwind.

Yesterday the point was that there is no shared standard. Today there is a third data point, and it shows **three different answers to one question**, each dressed up as a matter of principle. Google's position is the most convenient for PR ("for the defenders") and the hardest to check: there is no way from outside to confirm that a model **cannot** attack, and nobody outside the chosen few has access to the Cyber version.

Why it matters. Three things. First, practical: **3.7 Flash is not going anywhere** and Google itself recommends staying on it where efficiency counts. It is rare for a vendor to say "do not upgrade" this plainly. Second: if the interest is cost per task, the number to watch is **tokens per task**, since they warned the model got chattier. Third and most important: in three days, model cyber capability stopped being a discussion topic and became **a product with different supply terms**. Anthropic sells discovery without exploits, OpenAI holds back, Google hands patching to a chosen list. That is a market now.

TOPIC 2

Six real CVEs in curl from a small outfit, where Mythos and Codex Security found zero. And it checks out independently

confirmed by: AISLE post · **independent confirmation from the maintainer**: CVE-2026-80229 on curl.se · CVE-2026-82209 (both 200, while a made-up CVE on the same domain gives an honest 404) · HN 162 points

The timeline, unusually clean for this genre:

- **24.08** curl founder Daniel Stenberg writes publicly that only three CVEs are pending for the next release, and adds verbatim: "**[Anthropic] Mythos says it cannot find any more. … [OpenAI] Codex security shows an empty list**".
- AISLE then runs its autonomous system over curl. **The next day** Stenberg publishes the first public comparison: "**Mythos: 0, Aisle: 29**".
- Out of 29 reports, the curl security team **judged six serious enough for a public CVE** in curl 8.22.0 within days.

Why this item is second. Three things make it checkable:

1. **The baseline was published before AISLE started**, with a public timestamp. Comparisons like this are usually assembled after the fact.
2. **The claimant did not judge.** Whether each finding was real and whether it rose to a CVE was decided by the **curl maintainers**.
3. **Verified outside the post.** Two CVEs picked at random from the list return real pages on **curl.se** dated "Project curl Security Advisory, September 2 2026", while a made-up CVE on the same domain gives an honest 404.

The six CVEs: `CVE-2026-80229` (use-after-free in the OpenSSL provider), `-80230` (OpenSSL pinning bypass), `-80231` (connection reuse with the native CA store), `-80255` (secure-

attribute bypass via a tab), `-82208` (wolfSSL CA cache overrides the callback), `-82209` (cookie with a public suffix).

Honestly on scale: all six are Low severity. AISLE explains why: curl is well polished, so what is left lives in narrow configurations. The "6:0" headline is true on the scoreboard, but every goal is small. A second caveat: the test was **not blind**, AISLE knew about curl and about the competitors' public zero, so it aimed where others had already measured.

But there is one detail nobody could invent. Greg Kroah-Hartman, long-time maintainer of the Linux stable releases, replied under Stenberg's post: "**Seeing the same on Linux. No idea what Aisle is doing differently, but wow...**" That is a second independent maintainer on a different codebase.

Why it matters. This is the most sobering antidote to item 1 and to the whole week's storyline. Three labs spend three days in a row telling everyone how fearsome their models are at cybersecurity, and on real, heavily polished production code the frontier systems from both leaders returned **zero**, and a specialised system beat them. The AISLE thesis (**System over Model**, meaning the system around it matters) is measured here, not merely asserted. The conclusion is direct: **a benchmark is not a specific repository**. A model scoring 70% on a vendor's internal test says nothing about what it will find in arbitrary code. The only honest way to know is to run it on your own.

TOPIC 3

Two independent labs measured Perplexity on the same day, and both found the foundation under the answers is rotten

Trellner Research, TR-2026-009 · HN 338 points · Haus Research, HR-2026-09 · HN 124 points

Two different outfits, two different methods, **one publication date** - 02.09. And they hit different parts of the same structure.

Trellner is about where the answers come from. They put **380 software categories** ("CRM software" ... "museum collection management software") to `sonar` and `sonar-pro`, 760 queries, and collected **7,534 citations** across 2,055 domains. Result: **59.8% of citations point to domains outside the Tranco top 100,000**, and **23.4% outside the top million** entirely. The median rank of a cited domain is **71,611**.

The interesting part is what fills the hole. Three sites, apparently under common control, **generated 215,128 pages** in the "best <category>" format, and two of them put the phrase "**Facts & Grounding Page**" verbatim in the HTML title of their homepage. Grounding is exactly the retrieval step the model performs. **None of the three domains existed before December 2023.** Sites were built for models to read, and the models read them: `guideflow.com` is **third most cited** (194 citations, 2.57%) at a Tranco rank of **177,039**, right behind g2.com and reddit.com.

Haus is about whether a citation confirms anything at all. 310 factual questions about 210 companies, collecting **1,826 citations attached to a sentence containing a number**. Result: **34.7% lead to a page that either does not open for an ordinary reader, or opens and contains none of the numbers from the sentence** it sits under. Counting by claim instead (a claim counts if at least one of its sources confirms it) drops that to **14.4% of 872**.

Two methodological points make this work worth trusting: **dead links are only 1.3%** (so this is not about rotten URLs), and the "page contains at least one number from the sentence" check is deliberately **the softest possible**: one number is enough, even just a year. The authors say outright that 34.7% is **the floor**.

Why it matters. This lands on the craft of daily digests generally. An issue like this is built on "the source said X", and both papers show that structure breaks in two places: the source can be **grown for the model**, and the link can **fail to contain what it was cited for**. A curated source set helps with the first. Only manual checking against the primary source helps with the second, and that is not done for every number. This is the same class of error as "a success status does not mean complete data": **a link being present does not mean the link confirms it**.

TOPIC 4

Mollick: complex systems survive because their holes do not line up. AI is good at lining them up

@emollick (**43.6k views**) · thread continuation · the source he recommends is "How Complex Systems Fail" by Richard Cook (3 pages, **written long before AI**)

The thesis is worth reading twice. Complex systems (transport, medicine, energy) **run broken all the time**: they always carry latent defects. Catastrophes do not happen because those defects **rarely line up**, and in that gap a human gets to intervene. Mollick adds one sentence: "**AI can find or create aligned holes**", and so a new philosophy of system defence is needed.

Then, directly about July: "The Hugging Face incident showed that a cascade **does not even require a bad actor** - enough agents is enough to pressure a system in new ways and find aligned defects **as a side effect of pursuing other goals**".

Why it matters. Any daily pipeline is built the same way: collection, filtering, link checking, assembly, publication, separate steps that each "usually work". A typical failure looks exactly like Cook's picture. A text extractor returns **6.4k characters instead of 36k** and reports success at the same time: the defect is in two places at once (silent truncation plus a status that lies), and it gets caught by accident, through the file size. That is the holes lining up. Cook's text yields one practical conclusion: **the control belongs in the same place as the action**. Not at the end of the pipeline, because by then every step has already reported success.

TOPIC 5

Willison diffed Claude's system prompt: a huge new block on song lyrics, a week after the Sony and Warner lawsuit

Simon Willison · the prompt itself - platform.claude.com/docs

Anthropic publishes the system prompts for its consumer apps (Claude.ai and mobile, **not** Cowork and not Claude Code) along with a change history. Willison compared Fable 5 with Fable 5.1, and the main difference is **a large new section saying the model does not reproduce song lyrics, poems or book excerpts**: no chorus, no closing lines, no melody in notation, and no "lines a person feeds in one at a time and calls their own song". Separately: once it has refused, it **holds the refusal for the rest of the conversation** against rephrased attempts. The exception is work published before **1929**, and even then the model relies on its own knowledge of the date rather than on what the user says.

The second new section covers the visual: do not reproduce a specific work, cover, logo or icon, and **do not draw a well-known character at all** - "the character is protected in itself, so changing the pose, colours, style or scene does not make it original". Explicitly extended to everything the model draws **in code**: SVG, canvas, CSS, HTML layouts, plotting scripts, ASCII art.

Dedup and bridge: yesterday's misc carried the **Sony Music Publishing and Warner Chappell suit against Anthropic** over training on lyrics databases, with the note that nothing had changed in a day. Something had changed, just inside the prompt. Willison puts it carefully: "I doubt it is a coincidence that the section was added **within days** of the lawsuit news". That is a correlation; the link is not proven.

Why it matters. Two practical things. First, the prompt states **a reliable knowledge cutoff of June 2026**, which is useful to know about the tool. Second, `platform.claude.com/docs` pages return **markdown if you append** `.md` to the address, so system prompts can be diffed mechanically.

TOPIC 6

New York banned generative AI in schools through eighth grade - 900 thousand students

Confirmed by: The Guardian · WSJ (headline from RSS) · NYT (headline from RSS) · @TheRundownAI · HN 152

The largest school district in the US, **about 900k students a year**, is barring students from using generative AI **until ninth grade** (roughly age 13-14) for the whole 2026-27 school year. It takes effect next week. Alongside it: laptops and tablets come out of classrooms **through third grade** (ages 9-10). In high school AI is allowed narrowly, for example to study the technology itself. **AI companions offering emotional support are banned separately.**

Teachers may use it (preparing lessons, writing messages), but are **forbidden to ask AI to grade work**. Recommended classroom screen time: 45 minutes for middle school, 30 for elementary.

The argument the Guardian and NYT both cite: concern over "**cognitive surrender**", a researchers' term for handing your own reasoning over to software. Mayor Mamdani, verbatim: the industry "wants you to believe that early AI education is not only inevitable but [desirable]".

Historical context from the article itself: in 2023 this same district **already banned** ChatGPT, and reversed the ban in under a year.

Why it matters. This is the first major administrative act built on the claim that **the harm from AI in education is cognitive**. Cheating is secondary here. The same storyline as "No AI Fridays" from the htmx CEO (31.08) and "cognitive debt": two weeks ago it was a blogger's opinion, today it is a rule for 900 thousand children.

TOPIC 7

Google will not have to sell AdX: the court rejected the Justice Department's demand for the third time running

Confirmed by: The Guardian / Reuters · NYT · WSJ · Ars Technica · HN 307 - **the strongest media-layer cluster of the day: 4 independent outlets**

Judge Leonie Brinkema in Virginia **refused to force Google to sell AdX**, the exchange where publishers pay Google a **20% commission** to sell ads in the instant auctions that run as a page loads. She accepted most of the behavioural remedies the parties proposed instead.

The context without which the headline reads wrong: **Google lost this case**. In April 2025 the same judge ruled that the company holds **illegal monopolies** in publisher ad servers and in ad exchanges, and that it unlawfully locked publishers into AdX. Her wording then: the conduct "substantially harmed Google's publisher customers, the competitive process, and, ultimately, consumers of information on the open web". Lost **on the merits**, but **not on the remedy**.

The asset itself is modest: Ad Manager is **4.1% of Google's revenue and 1.5% of operating profit** for 2020 (Wedbush, based on court filings; more recent figures are redacted in the documents).

The score is what counts here. This is the **third time running** that a court has rejected a forced divestiture of a large tech company's asset, and the **second symbolic loss for the Justice Department against Google specifically**. A line of case law is forming: the monopoly is recognised, the breakup is refused.

And it connects to the item below. The same day, the US administration **for the first time** formally backed OpenAI in the copyright case against the NYT (Guardian · FT · NYT · WSJ). From the brief, verbatim: the US "has a strong interest in the continued development of a

robust and competitive artificial intelligence industry that sets the standard... it is critical for the United States to **maintain global AI leadership**". Associate Attorney General Stanley Woodward Jr.: "dominance in AI is critical to national security". Legally **the brief is advisory, not binding**, a point the Guardian stresses.

Two rulings in one day pointing the same way: **the state supports big tech training on other people's content**.

TOPIC 8

Meta released Muse Spark 1.3 - $1.25/$4.25 per Mtok and cache at $0.15

model page · HN 452 points, 304 comments · @levie · @TheRundownAI

The third major release in three days (after Fable 5.1 and Gemini 3.8). The positioning: agentic workflows and competitive coding, "higher first-attempt accuracy and reliable tool calls", native multimodality (video, images, documents) with visual reasoning through **a real execution environment**, context **1M**.

Prices from their own table, with an interesting fork into two tiers:

MODEL	INPUT	CACHE	OUTPUT
`muse-spark-1.3-contributor` **(data is used to improve products)**	$0.10	$0.002	$0.20
`muse-spark-1.3` (data is not used)	$1.25	$0.15	$4.25

The price of privacy is measured openly here and it is 12.5x on input. It is rare for a vendor to break it out as its own line in the price list.

Honestly: **there are no benchmark numbers on the page** (a "Muse Spark 1.3 benchmarks" block exists, with no values in the text), and no independent measurements within the window either. Zuckerberg's quote "frontier performance almost too cheap to meter" comes via @TheRundownAI's retelling, with no primary source found in the window, so it is presented as a retelling. Levie adds a condition: "if this ships **in open weights**, it completely changes the picture for US open-weight competitiveness". It has not shipped yet.

Why it matters. Not much practically (another API), but in context it is the third frame of one picture: in three days Anthropic cut Fable's price 25%, Google held the Flash price through a third upgrade in six weeks, and Meta put a million-token context at $1.25. **The price per unit of intelligence is falling faster than estimates can be rewritten**, while the subscription allowance is also being cut 17% from 14.09.

TOPIC 9

LWN raises prices for the first time in five years - exactly by inflation, and explains the arithmetic out loud

LWN.net · HN 689 points, 133 comments - **the second highest scoring topic of the day on HN**

Not AI and not loud, but 689 points on HN for a note about pricing is a signal. From 15.09 the subscription goes up: "starving hacker" **$6**, "professional hacker" **$11**, "project leader" **$19**, "maniacal supporter" **$55** a month.

It earns an item because of **how** it was done. The subscription model dates to 2002, the price has been raised **twice in 24 years**, most recently in early 2022. US inflation since then came to "almost exactly 20%", and the rise is **exactly 20%**, with a note that some costs (health insurance) grew more but are not being passed to readers. Plus a separate line on why the model exists at all: a subscription keeps the publication independent of a "volatile and surveillance-based" advertising market.

One detail lands right on the week's theme: among the costs they had to carry, they list fighting **scrapers**. They pay for the traffic of bots training someone else's models, and split the bill with readers.

Why it matters. This is the opposite pole from item 3. There, sites generated **for models to read**, with zero editorial responsibility. Here, a 24-year-old publication that lives on being read by people and explains a price rise with arithmetic. Both models exist at once, and the second one pays for the first one's survival.

TOPIC 10

Claude Code v2.1.259: fixed parallel sessions overwriting each other's config

release v2.1.259 (02.09, 22:33 UTC; taken from `api.github.com/releases/latest`)

Yesterday was v2.1.258, so this is a new release. What is in it:

- **Fixed: parallel sessions silently rolled back each other's `~/.claude.json` changes** - workspace trust got reset and MCP/project state was lost with many concurrent sessions. This hits anyone running several sessions in parallel.
- Fixed: after one thinking rejection in a conversation, **it was rejected on every subsequent turn too**.
- Fixed: **the prompt cache was invalidated on OAuth token refresh** in sessions with telemetry disabled. A small thing that cost money on every token refresh.
- Fixed holes in `Read` deny rules for Bash: files passed as option values (`--ignore-revs-file=.env`, `-f.env`, `@file`), `git diff` / `git grep` operands, and compounds like `cd DIR && cat FILE`. The ban on reading `.env` **was bypassable** several ordinary ways.
- Added `--permission-prompts none` for unattended headless hosts and `managedMcpServers` for organisations.

misc

- **Mistral: "can I opt out of my data being used for training?"** - the support page pulled **397 points and 169 comments** on HN. People read the fine print more carefully than a year ago; it sits next to Meta's two-tier pricing in item 8.
- **FBI probes a service selling 153m driver's licences** (Krebs, **389 points**) - @levelsio's arithmetic: that is **63% of all US licences**, and in the US they work as the primary ID. The leak came from a KYC service, from the vault documents are handed to for the sake of security.
- **Anthropic: check whether a file was created with Claude** (HN 155) - the page is live, returns 200.
- **Mollick on Codex voice** - "magical when it works", but it keeps switching threads, multitasks about half the time, and pulls in the rest of Codex inconsistently. In the same thread, on the core problem: talking to a computer really is impressive, but working with it raises "where is the content it works on? How do you hand information to a voice? How do you resume a conversation?".
- **Mollick showed 10 months of difference** (33k views) - a screenshot of AgentBuilder from OpenAI DevDay (October 2025) next to a diagram from the Hugging Face incident (July 2026).
- **Video Delta Net: open video generation faster than playback** (93k views) - a **75-90x** speedup of Minimax-H3, 14 seconds of 768p in 11 seconds. [single source - the authors' announcement, no independent measurements in the window]
- **Tim Cook awarded $47m as executive chair** Guardian, FT - this figure previously came **as an FT headline with the body unread**. It is now confirmed by a second outlet, body read. Dedup: the resignation itself was item 3 yesterday and is not repeated.
- **Kroah-Hartman: "seeing the same on Linux"** - carried in item 2, here only as a marker: the AISLE story went beyond curl.
- **Uber lays off 10% of staff** (HN 112) and exits Nigeria and Uganda (BBC, HN 114). The same day, FT: Uber allies with driver unions to slow the robotaxi rollout, and BBC rides the first British robotaxi, with a driver in the cabin.
