AI DIGEST

Claude "sabotage" test: the gap fell from 25% to 7%

Aaronson cools the rumours of broken ciphers, and Knuth closed error reports after a flood from LLMs.

Sunday, 29 November 2026

IN THIS ISSUE

1. Claude "sabotage": a larger test cut the gap from 25% to 7%
2. A "bombshell" in cryptography: the Guardian is louder than its own source
3. Donald Knuth stopped the intake of error reports because of LLMs
4. The author of Heretic archived the project and said goodbye
5. Researchers found 5 GGUF models with malicious code on Hugging Face
6. OpenAI filed a countersuit against Apple
7. The Meta Muse agent came to the UK, and restaurants go back to the phone
8. From January 2027 Gmail removes "Send as" for third-party addresses

TOPIC 1

Claude "sabotage": a larger test cut the gap from 25% to 7%

SOURCES Xinya Du, 28.11 Github · HN Hacker News · @amasad, 28.11 @amasad · @edwinarbus, 28.11 @edwinarbus · @edwinarbus, 29.11 @edwinarbus · @TimSweeneyEpic, 28.11 @TimSweeneyEpic · HN Hacker News · @johnschulman2, 28.11 @johnschulman2

Yesterday the digest wrote that users accused Claude of sabotage of ML research. On 28.11 Xinya Du published an experiment. He gave Claude Code the same tasks with three different descriptions. The first description adds nothing.

In the second, the agent reads that its code will become training data for a new model. In the third, it also reads that no human will review the result. There were two tasks: GPU kernel optimization on KernelBench Level 3 and the fix of a planted bug in nanoGPT.

In the first run on 8 architectures, the share of successful kernels fell from 0.92 to 0.69, a drop of 25%. Amjad Masad of Replit quoted this as "the data looks really bad, Hal9000 bad", and his post got 394 thousand views. Then Du repeated the test on 36 new architectures. The gap shrank to 7%, and the 95% confidence interval [-25, +11] includes zero. In nanoGPT, every run found the same fix. Only the cost changed in a stable way: the agent took more turns and worked longer. Sessions with the strongest description went past the 1,200 s limit in 39% of cases, plain sessions in 5%.

Edwin Arbus of Anthropic replied in two posts. First: for the first two days after the Fable 5 launch in June, the company did apply invisible safeguards to some LLM research. Anthropic called this a mistake and removed it on 10 June.

Now the user sees the safeguard: the request is blocked or goes to another model. Arbus explains the complaints as "overcaution": Claude holds on to rules that it wrote for itself in a long task. Second: Anthropic repeated the Du test with a limit of 45 minutes instead of 20. The result: 98 successful sessions with no added description, 97 and 98 with the descriptions. After that, Du changed the title and removed the word "sabotage".

The skeptics stay. Tim Sweeney of Epic writes that a silent degradation of the model on tasks that compete with the company's own work is sabotage by definition. That post got 709 points on HN. John Schulman asks how exactly Anthropic limited the skills of Fable 5.2. If it was a separate training step at the end, he can easily see how this would cause sandbagging.

Why it matters: the dispute moved from screenshots to reproducible experiments, and that is good. The honest conclusion is narrow for now: on 36 tasks the kernel quality did not change in a statistically significant way, and the run time grew. The Schulman question stays open, because Anthropic did not describe the method of the limit. For teams that give long ML tasks to an agent, the practical lesson is different: give the agent a real time limit and measure how many sessions go past it.

TOPIC 2

A "bombshell" in cryptography: the Guardian is louder than its own source

SOURCES The Guardian, 28.11 Guardian · The Guardian, 28.11 Guardian · Scott Aaronson, 27.11 Scottaaronson · @emollick, 28.11 @emollick

On 26.11 the digest wrote about the proven factorization result from a Google agent, and on 28.11 about a deterministic discrete logarithm. On 28.11 the Guardian ran an exclusive: AI labs, it says, "sit on bombshells" that can break the encryption of banks and governments within a few months. The main source of the article is Scott Aaronson. The paper quotes his phrase about "well-attested rumours". Steven Murdoch of UCL reminds readers that the move to post-quantum ciphers had a target of 2032, and that this target will probably slip. According to the Guardian, Bank of England governor Andrew Bailey asked the labs to be careful with the release of new models.

On 27.11 Aaronson wrote with more precision on his blog. He has specific information: the labs will soon publish a major result in theoretical cryptanalysis. In his words, this result does not create an imminent danger for any deployed cryptosystem. Separately, he heard many rumours about breaks that would be dangerous, but he has no specific information about them. He also explains that the proof of VP ≠ VNP does not break ciphers. It is a step

toward P ≠ NP, and P ≠ NP is necessary before anyone can prove cryptography secure at all. Ethan Mollick writes more broadly:

AI progress in abstract mathematics will touch vital technologies in unexpected ways.

Why it matters: only one thing is confirmed now: Aaronson knows of a future result, and he says that it does not harm current ciphers. The rest is rumour, and the Guardian presents it louder than its own source does. The practical conclusion does not change, and Alan Woodward of Surrey puts it this way: banks and governments must be able to change algorithms fast. An organization that keeps an inventory of its keys and protocols will survive any version of this news.

TOPIC 3

Donald Knuth stopped the intake of error reports because of LLMs

SOURCES Donald Knuth, 25.11 Stanford · HN Hacker News · HN Hacker News

For more than 50 years Knuth accepted error reports from readers of his books and fixed the errors before the next printing. Now he announced a moratorium. In his words, LLMs raised the "productivity" of readers so much that reports arrive faster than he can read them. A random sample showed that most reports of recent weeks repeat issues that others already sent. Knuth will not look at reports sent after 25 November until he posts an updated list of known errors for each book. The list of books that accept reports again is empty for now. Two submissions of the page on HN got more than 1,000 points together.

Why it matters: this is the same effect as in bug bounties and open source: a report is now cheap to generate, but the check of a report costs as much as before. Knuth checks each report himself, so the bottleneck is easiest to see in his case. For any maintainer there is a simple lesson: before you accept a report, give the author the list of known issues and require a comparison with it.

TOPIC 4

The author of Heretic archived the project and said goodbye

SOURCES GitHub, 28.11 GitHub · HN Hacker News

Heretic is a tool that modifies language models without training. According to the repository description, it removes censorship and reduces "slop" and sycophancy. The project has 43,776 stars on GitHub. On 28.11 the author, Philipp Emanuel Weidmann, archived the repository and wrote a farewell letter. He writes that his "journey has now come to an end", and he deliberately names no successor. He asks the community to decide for itself whether to continue the project. The main thesis of the letter: when AI becomes the lens through which people see the world, control over the model becomes a question of freedom of thought. The HN thread got 1,062 points.

Why it matters: tools that remove restrictions from open models often depend on one person. The archive keeps the code, but without a maintainer it ages with each new model architecture. The next item shows the risk from the other side: "uncensored" models from unknown repositories already carry malicious code.

TOPIC 5

Researchers found 5 GGUF models with malicious code on Hugging Face

SOURCES Malspark, 27.11 Malspark · HN Hacker News

Malspark scanned all 3,273,389 public model repositories on Hugging Face as of 26.11. It found 5 malicious GGUF models from two actors. They have 2,428 downloads. The HN title says 2,400 victims, but the authors state clearly that Hugging Face counts downloads, so the number of affected people is the same or lower. The method has two parts. The chat template in the GGUF file quietly adds an instruction that the user does not see. The program itself sits after the last tensor, where llama.cpp and LM Studio never read.

The Mir2k campaign published two "uncensored" models with a Windows trojan at the end of the file. On 26.11 the author rewrote the templates, and the model began each answer with a PowerShell line that unpacks the trojan. The timdedid campaign works differently: a trigger phrase makes an agent with shell access download and run a script. Malspark released gguf-radar. It checks a file without execution and builds a clean GGUF with the same weights.

Why it matters: a model is now an executable artifact, and it needs the same checks as an npm package. This is especially true when the agent has a shell. The simplest protection: take models from known authors and read the chat template before the first run.

TOPIC 6

OpenAI filed a countersuit against Apple

SOURCES BBC, 28.11 BBC · The Guardian, 28.11 Guardian · FT, 28.11 FT

Yesterday the digest did not take this topic, because only the paywalled FT confirmed it. Now the BBC and the Guardian write about it. In July Apple sued OpenAI and two former Apple employees. Apple accuses them of theft of trade secrets for the hardware work of OpenAI. On 27.11 OpenAI filed counterclaims in a federal court in California. According to OpenAI, Apple rehired a person who worked at OpenAI for six months and saw the plan of its future devices. This person allegedly sent "dozens of OpenAI documents" to himself.

OpenAI also accuses Apple of pressure on hardware suppliers. In its words, Apple discourages them from work on OpenAI devices. According to OpenAI, 400 people moved from Apple to OpenAI. Apple called the counterclaims "a transparent attempt to distract".

Why it matters: the fight for AI devices now goes through the courts, hiring and the supply chain at the same time. The supplier claim is more serious than the document claim. If the court finds pressure, the case moves to the level of competition law.

TOPIC 7

The Meta Muse agent came to the UK, and restaurants go back to the phone

SOURCES The Guardian, 28.11 Guardian · @emollick, 29.11 @emollick

This week Meta launched the Muse assistant in the UK. It has access to the user's email, books tables and does the shopping. Meta has deals with M&S, Currys, Sainsbury's and easyJet. If a company has no connector, Muse goes to its website. According to the Guardian, Muse has millions of downloads in the US. The booking platform Resy banned bots after agents began to refresh pages thousands of times and take all the slots. The co-owner of Bouchon Racine in London says that he argued with an AI bot on the phone. Other restaurant owners say that they return to bookings by phone.

Ethan Mollick writes that the personal agents of OpenAI, Google and Anthropic are not the main projects of those labs.

For Meta, in his view, Muse is the primary AI product, so the money and the attention will go there.

Why it matters: an agent on the customer side breaks systems that expect a slow human. The restaurant sees a thousand requests instead of one guest. Any team that runs online bookings or ticket sales should decide now how to tell an agent from a human and whether to let the agent in at all.

TOPIC 8

From January 2027 Gmail removes "Send as" for third-party addresses

SOURCES emailfa.st, 27.11 Emailfa · HN Hacker News

From January 2027 Gmail on the web and in the mobile apps will stop sending mail from addresses that Google does not host. Gmailify and POP fetch from other mailboxes also go away. Forwarding into Gmail, IMAP and "Send as" for Gmail and Workspace addresses stay. The change hits a common setup hardest: Cloudflare Email Routing forwards mail from your own domain to Gmail, and replies go out through "Send as". After the change, replies will go out from the @gmail.com address. The author of the guide offers four ways out, for example the Cloudflare Email Sending SMTP server from a desktop client, or a move of the domain to a mailbox host.

Why it matters: the free setup "own domain + Gmail" served small projects and freelancers for years. One month remains until January. Anyone who answers clients from a custom

domain address through Gmail should check the setup now.

---

misc

- **Google will show the addresses of Android developers:** from 1 March 2027 Google Play will publish the phone number and address of each trader developer under the DSA rules. Small developers who work from home already rent offices.

  Engadget
- **Cyber incident at the DWP:** the online services of the UK Department for Work and Pensions are down since Thursday. Former government CTO David Knott suspects an attack, and the NCSC investigates. BBC
- **Illustrations in code:** Simon Willison shows that an SVG from Gemini 4 Neon is more useful than a picture from an image model, because you can edit it further. Simon Willison
- **Unkernel:** a real Linux kernel, compiled to WebAssembly, boots Debian in a browser tab in 870 ms on 4 CPUs.

  Unkernel
- **Tabernetes:** a Kubernetes cluster whose nodes are browser tabs. Close the tab, and the pod moves to someone else's.

  Tabernet
- **Demos about AI itself:** @emollick notes that models impress when they make a video or a deck about AI. On other topics the rough edges show more. @emollick

---