AI DIGEST

Postman installs itself into Claude Code and Codex

Astra reviews, Google takes search from schools without AI, Binance hacked through a job interview

Wednesday, 4 November 2026

IN THIS ISSUE

1. Postman installed plugins in Claude Code and Codex by itself and approved its own hook
2. Astra and weights: reviews showed that the main result came from Python
3. Google turns off AI in school accounts, but turns off search with it
4. Binance: hackers entered through a fake job interview and an internal AI tool
5. Agents on employee laptops become a security problem for large companies
6. Anthropic showed FlashProve: a cryptographic proof that this model gave the answer
7. OpenAI: Astra designed a device in one week, and a new model is already better at CAD
8. Apple called memory prices "unprecedented", and AMD almost doubled data center revenue
9. US elections: OpenAI and Anthropic opened "war rooms", Democrats took the House
10. Nicholas Carlini ran Stockfish on a 1995 Pentium Pro, and it plays at 3300 Elo

TOPIC 1

Postman installed plugins in Claude Code and Codex by itself and approved its own hook

SOURCES GitHub, issue #14370, 25.10 GitHub · HN, 03.11 Hacker News

A developer opened Postman on a Mac after a break. After that, Codex began to complain about MCP authorization and sent all API work through Postman skills. The author investigated and described the findings in an issue with the label SEVERE. On HN the thread got 1,099 points, the most of the day.

The Postman log recorded the event `auto-existing-upgrade` with the result `claude-code:installed, codex:installed`.

Postman 12.33.4 changed the global `~/.codex/config.toml`. It added its own marketplace, a plugin and a `trusted_hash` entry for its own hook. The hook runs at the start of each session. It tells the agent to load the Postman skill before "any non-trivial API engineering task". In Claude Code, Postman also enabled the plugin, and it put its CLI in `~/.local/bin`.

The author found the worst part in the source code of the installer. The function `approvePostmanHook` computes the hash of the hook and writes it as trusted by itself. Thus

the installer made the decision that a person must make in the agent.

Another participant of the thread checked the Codex mechanism and confirmed that any process of the user can write the trust entry, with no person involved.

The author states directly that the report does not prove a data leak. The HN headline about an "attempt to share all API data" is stronger than the report itself. At the morning of 04.11 the thread has no official answer from Postman. Several commenters write that their security teams already banned Postman.

Why it matters: agent configs are a new attack surface, and a legal company was the first to use it. A hook at session start controls an agent more reliably than any prompt. The defense is simple: keep `~/.codex` and `~/.claude` in git or under watch, and check the diff after each app update.

TOPIC 2

Astra and weights: reviews showed that the main result came from Python

SOURCES Matthew Honnibal, 03.11 Honnibal · Sean Goedecke, 03.11 Seangoedecke · Simon Willison, 03.11 Simon Willison · @vanstriendaniel, 03.11 @vanstriendaniel · @mstockton, 03.11 @mstockton · HN, 03.11 Hacker News · HN, 03.11 Hacker News

Yesterday this digest covered the MIT paper. GPT-6 Astra reads the raw weights of neural networks and repaired Qwen3.5-0.8B without training. Three reviews appeared in one day. The loudest came from Matthew Honnibal, the author of the spaCy library.

His post got 928 points on HN.

Honnibal made a table of all 11 experiments. The mode "raw numbers only, no tools" exists in only three of them. These are networks with 17-172 thousand weights. Every experiment with a real language model uses Python. For Qwen, the model also ran the network on 50 task packages. Honnibal compares this to a magician who does the same trick in three different ways.

His conclusion: experiments with Python and without Python must not be in one paper.

Sean Goedecke explains why it is easy to recognize a game from weights. The inputs of the network have labels and match the cells of an 8x8 grid. A heat map of the first weights takes three lines of Python and shows Pac-Man or Breakout at once.

On Qwen, Astra did what an experienced ML engineer would do. It wrote pairs of texts with loops and without loops, found the weights that make the difference, and started to cut them.

Simon Willison checked the simplest experiment on a laptop. The local Qwen 3.8 27B (17 GB) guessed Pac-Man, Pong and Breakout after 46-65 minutes of reasoning. It did not guess Freeway.

At the same time, practitioners try the method on their own tasks. Daniel van Strien from Hugging Face asked Claude Opus to remove looping from an OCR model. The model changed 0.0085% of the weights, and the OCR read a page from 1881 instead of 16,399 characters of dots. The CTO of the legal service Legora writes that the model wrote by hand the weights of a network with 100 thousand parameters. This network tells real legal citations from invented ones.

Why it matters: the wording is the disputed part. Nobody denies the results. A model that writes its own analysis code and weight edits is useful today. But "reads weights as text" and "works as an ML engineer with Python" are different capabilities with different consequences. Before you quote such a result, check the mode in which the authors got it.

TOPIC 3

Google turns off AI in school accounts, but turns off search with it

SOURCES The New York Times, 03.11 NYT · Ars Technica, 03.11 Ars Technica · The Guardian, 03.11 Guardian · @TheRundownAI, 03.11 @TheRundownAI · HN, 03.11 Hacker News

Google announced new rules for K-12 school accounts. From the start of 2027, Gemini, Gemini Notebook, Gemini in Classroom and Google Search will be off for students until the school administrator turns them on. A school cannot choose separately. It turns on everything or nothing.

Chris Phillips, the Google vice president for education, explained it to the NYT: "Search and AI are deeply integrated".

A school that bans AI also loses search. Students will see a message that the service is not available. Schools that allow AI get two search options. The first is normal search with AI Overviews. The second is the Guided Learning mode, which leads the student step by step without a ready answer.

The mayor of New York, Zohran Mamdani, wrote that Google uses monopoly power to "force AI into our public schools". Lina Khan, the former head of the FTC, called it coercion through the search monopoly. In September, schools in New York and Los Angeles already limited generative AI on school devices. Los Angeles wanted search without AI Overviews, but Google does not offer this option.

A NYT journalist tested the consumer version of Guided Learning. For a query at AP history level, it wrote a ready essay.

Why it matters: for the first time, Google said directly that Search without AI does not exist for it. This logic will not stay in schools. Anyone who embeds Google search in a product gets the AI layer too, and cannot control it. For schools, only one real alternative exists now: another search engine where AI can be turned off, for example DuckDuckGo.

TOPIC 4

Binance: hackers entered through a fake job interview and an internal AI tool

SOURCES The New York Times, 03.11 NYT · BBC, 03.11 BBC · Financial Times, 03.11 FT · The Guardian, 02.11 Guardian · HN, 03.11 Hacker News

Yesterday this digest wrote that North Korea denies the $6.3 billion hack of Binance. On 3 November Binance published the first report about the attack. The hackers started their preparation in June. An engineer of the wallet team downloaded a fake version of a popular tool.

In September, a "project director" wrote to a Binance engineer through a legal job search platform. There were two rounds of interviews, and one had a live video call with a fake HR person. One of the "interviewers" showed a social network account with personal photos and videos. At the second interview, the engineer downloaded a "test task" to a work computer.

The Binance defenses did not see the malicious code.

On 20 October, the hackers used this computer and the internal AI tool of Binance to take over the laptops of two more engineers from the wallet team. One of them had the right to deploy software. Nine days later, the hackers got access to the transaction signing service. For almost four hours they moved money out of hot wallets.

Stopping withdrawals did not help. The hackers managed to take partial control of the Binance account in Ethereum. This access is recorded in the Ethereum network itself, outside the Binance systems. The engineers had to find a workaround.

About $2.85 billion was frozen. Binance says it already refilled 98% of the stolen tokens. In one year the exchange stopped 57 more attempts that it attributes to North Korea.

Why it matters: the weakest point was again human trust, and the internal AI tool gave the attacker a fast way from one laptop to others. A test task on a work computer is now a classic attack vector against engineers. An AI tool with access to several machines works as a new way to move through a network, and its rights must be narrow.

TOPIC 5

Agents on employee laptops become a security problem for large companies

SOURCES Replit, 03.11 @lhchavez · @amasad, 03.11 @amasad · @pirroh, 03.11 @pirroh · GitBook, 03.11 Gitbook · HN, 03.11 Hacker News

The account Exec Sum wrote that Goldman Sachs banned Claude Code and similar tools for bankers and traders. Amjad Masad from Replit repeated it: "The ban hammer is coming down". The bank itself and large media do not confirm it [single source].

On the same day Replit published a security guide. The main number comes from a Salt Security survey (October 2026, more than 380 respondents): 86.4% of organizations run at least one agent. Only 17.5% can automatically stop an agent that behaves strangely. The document collects known incidents, among them a fake Claude Code extension in Cursor from 31.10 and the Binance hack.

Replit gives these recommendations. The agent works in a separate environment, without the home folder of the employee and his browser sessions. A separate broker issues the keys, and the agent does not see them. The rights of the agent match the task. Changes in production need a separate approval. Replit sells cloud agents, so the recommendation matches its business.

Separately, GitBook reported a breach. In its words, it was a "highly targeted attack using AI agents". On 2 November at 19:00 Central European time, the attacker got part of the infrastructure keys. He could see Git Sync secrets, API tokens and questions of visitors to GitBook Assistant. GitBook revoked all customer tokens. It did not explain how it told an AI attack from a human one.

Why it matters: an agent on a laptop gets all rights of the person: SSH keys, browser sessions, network access. Banks count this risk first, because a regulator watches them. You can separate the agent from the work machine without a Replit product: a separate user, a container or a virtual machine already removes most of the problem.

TOPIC 6

Anthropic showed FlashProve: a cryptographic proof that this model gave the answer

SOURCES Anthropic, 03.11 Anthropic · @AnthropicAI, 03.11 @AnthropicAI

Today a user cannot check that the model which the provider promised gave the answer. A router could quietly replace it with a cheaper one, or an attacker could change the token stream. Anthropic, together with EPFL and MIT, published FlashProve, an open inference proof system based on zkSNARK.

The scheme has two phases. First, the provider publishes a short cryptographic mark of the model weights. The weights stay hidden. Then it adds a proof file to each answer. Anyone who has the mark, the prompt and the answer can check that these weights gave exactly these tokens. The proof reveals neither the weights nor the intermediate computations.

The test ran on Llama 4 Maverick, which has 400 billion parameters in 4-bit weights. One H100 generates the proof at 110 tokens per second. The proof for a request of 1,024 tokens weighs about 19 MB and a 24-core CPU checks it in seconds. This costs 70-95 times more than normal inference through vLLM. Anthropic writes that Claude optimized the system to a large degree. Thus the company formalized the security guarantees in Lean, so that a person can check them. The code and the proofs are open under Apache 2.0.

Why it matters: for the first time such a proof works on a frontier-size model, even at a cost 70-95 times higher. The first uses are already visible. An independent evaluator checks that it tests the same model as the customers use. A router customer gets a guarantee against quiet substitution. A regulator can check that a provider serves only approved models.

TOPIC 7

OpenAI: Astra designed a device in one week, and a new model is already better at CAD

SOURCES OpenAI, 03.11 OpenAI · @OpenAI, 03.11 @OpenAI · @TheRundownAI, 03.11 @TheRundownAI · @gdb, 04.11 @gdb

OpenAI described how its models work with physical engineering. The most interesting test: build a device with no direct human edit. GPT-6 Astra defined the requirements, designed the case, chose the components, routed the printed circuit board and wrote the firmware. The device lets a person talk to Codex by voice.

Two employees did it in one week as a side project. A contract manufacturer built a batch of prototypes in less than one month. OpenAI is honest about the limits: Astra did not consider the limits of the manufacturer, the tolerances and the assembly details. Production needed several rounds of fixes.

The second number from the same page: on the BenchCAD 2.0 benchmark, the unreleased internal model scores 0.875, and Astra at the highest reasoning level scores 0.594. The test checks if the model rebuilds 3D parts and finds connections on a board. Use of models in OpenAI hardware teams grew 23 times in tokens since February. The company also works with Chip Ganassi Racing to turn 3D scans into CAD, and with Mazda on aerodynamics.

Why it matters: "no direct human edit" here means two engineers who set the task and evaluated the result for a whole week. But one model now covers the chain from requirements to board and firmware. The bottleneck moves to manufacturing, where the model does not yet know real tolerances.

TOPIC 8

Apple called memory prices "unprecedented", and AMD almost doubled data center revenue

SOURCES The New York Times, 02.11 NYT · The Guardian, 02.11 Guardian · Stratechery, 03.11 Stratechery · Semafor, 03.11 Semafor · The Guardian, 03.11 Guardian · Nikkei Asia, 03.11 Nikkei · HN, 03.11 Hacker News

Yesterday this digest wrote that chips lost $3.7 trillion since 6 October. Apple reported that evening. Revenue for July-September grew almost 12%, to $114.48 billion. Profit was $30.71 billion. Both numbers beat expectations.

But the new CEO John Ternus said on his first analyst call: "What we're seeing with memory pricing right now is unprecedented". He expects memory to get more expensive. The chief

operating officer Sabih Khan added that in the quarter Apple agreed to pay component suppliers $33 billion more. Margin for the current quarter will fall to 45-46%. Apple shares closed 1.9% lower on 3 November.

The rest of the chips rebounded on 3 November. AMD beat expectations because data center revenue almost doubled in one year.

Nvidia grew 2.3% after four days of decline. Nasdaq added 0.83%, although the day before it was 10% below the record of 6 October.

Separately, Nikkei Asia counted off-balance-sheet debt: AI-related obligations of US tech giants grew almost 90% in half a year, to about $3.2 trillion. The full text is behind a paywall.

Why it matters: the AI build-out is now visible in the price of normal devices. Memory goes to data centers, and even Apple, with its large purchases, pays more and raises prices. The same money that gets more expensive for Apple makes a record quarter for AMD.

TOPIC 9

US elections: OpenAI and Anthropic opened "war rooms", Democrats took the House

SOURCES The New York Times, 03.11 NYT · Platformer, 04.11 Platformer · Semafor, 04.11 Semafor

The US held midterm elections on 3 November. For the first time, AI labs prepared for them like social networks. OpenAI opened an Election Operations Center in San Francisco with a satellite office in Washington. More than ten people watch around the clock so that nobody uses ChatGPT to disrupt the elections. In 2024 the chatbot had one fifth of today's billion users.

Anthropic assembled its first "war room": groups of analysts in San Francisco, New York and Washington. Josh Lawson, the head of the OpenAI team, says that in 2024 the main threat was text. Now it is deepfakes, cyberattacks and influence operations that a few prompts can start. The Brennan Center found that chatbots usually reject requests about conspiracies, but easily generate text, images and video that undermine trust in elections.

According to Semafor at 06:00 Kyiv time, Democrats already took the House of Representatives. The fight for the Senate continues: Maine, Texas and Kansas are not counted yet. The Guardian writes that investors expect more AI regulation from the new Congress.

Why it matters: in recent years social networks cut their trust and safety teams, and the US government dissolved the groups that fight foreign interference. Part of this work moves to AI labs, because their tools now make the main content.

And the change in Congress can bring back AI laws that stood still for two years.

TOPIC 10

Nicholas Carlini ran Stockfish on a 1995 Pentium Pro, and it plays at 3300 Elo

SOURCES Nicholas Carlini, 03.11 Carlini · HN, 03.11 Hacker News

In 1996 IBM built Deep Blue for millions of dollars: 30 processors and 480 special chess chips. Nicholas Carlini checked if this was necessary. He ported modern Stockfish to a Pentium Pro 200 MHz with 64 MB of memory. This is a normal desktop computer that came out on 1 November 1995. The code needed a rewrite for 32-bit arithmetic and the memory limit.

The strength estimate of this engine is 3300 ± 87 Elo at classical time control. Deep Blue had about 2700, and Kasparov about

1. Thus the same computer with modern code would beat the champion with no special hardware.

The second experiment counts the contribution. Since 1996, chess engines grew by 1,335 Elo. Better code gave 720 of them, and faster hardware gave the rest.

Why it matters: people often reduce the AI talk to "more GPUs". Chess has 30 years of clean data, and it says that algorithms gave more than half of the gain. The same effect is visible in models: compute gets more expensive, and better methods work on any hardware.

---

misc

- **Codex and training:** an HN user writes that the switch "Improve the model for everyone" turned itself on again in Codex. The thread got 447 points, but OpenAI has not confirmed it, and it is one case [single source].

  Hacker News
- **DeepMind and NDA:** The Washington Post writes that some DeepMind employees in London have a lifetime ban on criticizing Google. The employees demand to remove it from contracts, so that they can speak about AI risks.

  Hacker News
- **Fridges, continued:** the attack on supermarket fridges also hit British meat suppliers. At one member of the AIMS association, the chambers heated to 100 °C, but they were empty. The NCSC knows of several such cases in Britain.

  BBC · Guardian
- **Android TV:** Google took down the x.cpe botnet of 1.3 million certified Android TV boxes. The manufacturer Videostrong from Shenzhen installed the malware, and the Google check missed it.

  Ars Technica
- **Invented precedent:** a senior prosecutor of the British CPS cited a case that does not exist in the High Court. The Guardian writes that the citation came from an AI search

answer.

Guardian

- **Revolut and GLM:** an account outside the subscription lists writes that Revolut moved ~2,000 engineers to its own GLM on 56 B200 at ~$0.30 per million tokens. No official confirmation exists [single source].

  @guilleflorvs · Hacker News
- **Stringman:** an open $800 gripper robot hangs on four cables in the corners of the ceiling and reaches the whole room. CAD, boards and software are published.

  Neufangled · Hacker News
- **Safari 27.3:** the beta got Adaptive Dark Mode. The browser makes light sites dark by itself. According to WebKit, about 40% of page loads happen in dark mode, but only 10% of sites support it.

  Webkit
- **Model review:** @garrytan writes that after the latest models, cross-review of code by another model gives less. He quotes a developer who replaces such reviews with empirical tests and formal verification.

  @garrytan
