AI DIGEST

Nvidia board probes Jensen Huang over Chinese spy tie

Also: harness evals for Opus 5.5, Goldman keeps Claude Code for engineers, 284B DeepSeek with no GPU.

Monday, 9 November 2026

IN THIS ISSUE

1. Huang admitted the relationship, and the Nvidia board started an independent investigation
2. Open evals: the harness hardly changes the result of Opus 5.5, but it changes the price
3. Goldman Sachs kept Claude Code for engineers only
4. DeepSeek-V4-Flash with 284B parameters runs on a PC with 64 GB of memory and no graphics card
5. A license plate opens a car: GPT 5.6 helped the researchers
6. Texts from AI clog the work of public bodies
7. Science: EEG headsets feed brain models with user data
8. Outlook.com started to reject some mail from domains without DMARC
9. Collison: nobody has yet described a positive AI scenario
10. Agent memory: nine formats, and none reads the others

TOPIC 1

Huang admitted the relationship, and the Nvidia board started an independent investigation

SOURCES Nvidia, 08.11 Nvidia · Axios, 08.11 Axios · The New York Times, 08.11 NYT · The New York Times, 08.11 NYT · Semafor, 08.11 Semafor · HN, 08.11 Hacker News

Yesterday's issue said that the FT saw a US intelligence memo about the relationship of Huang with Alice Huang of the Singapore company Megaspeed. On Sunday, the story moved from anonymous sources to statements by the people involved.

Huang sent a letter to Nvidia employees. "I was unfaithful to my wife", he writes, and he apologizes to his family, colleagues and shareholders. He promises to cooperate fully with the investigation. Axios published the text of the letter. Semafor and the NYT quote the same phrases.

The board of directors announced an independent investigation. A committee of independent directors leads it with an outside lawyer. The board writes that it has no reason to believe that national security or Nvidia secrets suffered.

Spokesman John Rizzo said that Huang does not take part in the choice of the lawyer and stays CEO "without interruption".

The NYT adds two details. This spring, Secretary of State Marco Rubio led the preparation of a secret memo about this relationship. Officials got the memo before the May trip of Trump to China. On Sunday, Trump said that he "has not heard that Huang did something bad" and wants him to stay CEO.

Why it matters: Huang runs Nvidia directly, and he has about 60 direct reports. Thus the investigation by the board checks if a company worth more than $4 trillion can control its founder. Semafor named it this way: a test of "founder immunity".

TOPIC 2

Open evals: the harness hardly changes the result of Opus 5.5, but it changes the price

SOURCES g-harness-evals, 08.11 GitHub · @garrytan, 08.11 @garrytan · caleb.software, 21.09, updated 07.11 Caleb · HN, 08.11 Hacker News

Garry Tan opened a repository with evals of agent harnesses. The model and the tasks stay fixed. Only the harness changes: the system prompt, the tools and the memory. There are 20 tasks, and they are public.

The results on Opus 5.5 are as follows. Claude Code solved 76.3% of the tasks at $0.22 for each solved task. Hermes Agent scored 78.8% at $0.24. The Pi loop scored 72.5% at $0.19. OpenClaw scored 76.3%, but at $0.44 and 2.5 times slower. The authors write directly that the difference in the share of solved tasks between harnesses is not visible. A gap of up to 7.7-14.2 points can hide in a sample of this size.

The most honest finding concerns Tan himself. Opus 5.5 loaded his skill set gstack every time, but did not run a single step: 0 of 46 sessions. At the same time, the price of a run grew about two times. In Claude Code, a rule in the system prompt worked in 30 of 30 turns. Without the rule, it worked in 18 of 30. The model did not open a skill that the prompt does not name in any of 30 turns.

The post "Goodbye Claude, Hello Pi" got 713 points on HN. The author moved from Claude Code to the minimal harness Pi.

His argument: in one year Claude Code had 244 releases, and more than 3/4 of them changed the system prompt or the tool descriptions. The description of one Bash tool takes almost 3,000 tokens. The comments remind readers of the price of minimalism: Pi has no mode that blocks dangerous actions of the agent.

Why it matters: with a strong model, the harness decides the price, the speed and the control, and the model decides the share of solved tasks. A skill works only when the prompt names it directly. Check this with an eval on your own tasks, because 20 public tasks give wide intervals.

TOPIC 3

Goldman Sachs kept Claude Code for engineers only

SOURCES The State of AI, quoting Bloomberg, 02.11 Thestateofai · HN, 08.11 Hacker News [single source]

On 04.11, this issue had only a rumor from one tweet: Goldman bans Claude Code for bankers. Now there are details that Bloomberg got from sources in the bank. The Bloomberg article is behind a paywall. Thus the facts below come from the summary by The State of AI.

In October, the bank updated its internal rules. Claude Code and GitHub tools are now open only to engineers. There are about 12,000 of them among about 47,000 employees. Strats, who work between trading and development, kept access.

Traders wrote scripts with agents to track client flows and position models, and some of them are not happy.

In July, CIO Marco Argenti praised "vibe coding" by non-technical staff. In April, the bank already removed Claude for bankers in Hong Kong. At the same time, Goldman and Morgan Stanley lead the IPO of Anthropic.

On HN, the most popular explanation is technical. An agent on the machine of a trader has client data, email, the internet and the internal network at the same time. This is the "lethal trifecta": such an agent can read a secret, get a malicious instruction and send the data out.

Why it matters: the code of an engineer goes through review and tests. The script of a trader often goes into daily work without them. In a regulated industry, the bank chose a simple rule: only a person who can read the code gets an agent that writes and runs code.

TOPIC 4

DeepSeek-V4-Flash with 284B parameters runs on a PC with 64 GB of memory and no graphics card

SOURCES llama-modchip, 08.11 GitHub · HN, 08.11 Hacker News

The project llama-modchip is a set of patches for llama.cpp, a quantizer and a wrapper to run the model. The main demo is DeepSeek-V4-Flash with 284.33 billion parameters, and 13.17 billion of them are active. The model weighs 56.56 GiB, which is 1.71 bits per weight on average. The normal build UD-IQ1_S from Unsloth weighs 76.87 GiB and does not fit in 64 GB of memory.

The authors compressed the model like this. The 256 experts of each layer are mostly in IQ1_S (1.56 bits). The shared layers and the attention stay at 4-5 bits. With a 256K context, the process takes 57.61 GiB of memory. On a 16-core Zen 5, prompt processing runs at 44.4 tokens per second, and generation runs at 9.1-14.7 tokens per second. Normal llama.cpp, Ollama and LM Studio cannot open such a file, because it uses its own tensor type.

The criticism on HN concerns quality. The README compares the patched engine with the unpatched engine on the same compressed model. The README has no comparison with

the full DeepSeek-V4-Flash. Commenters remind readers that below 4 bits, quality usually falls in a non-linear way.

Why it matters: a frontier-class model now fits in a desktop PC, and for private data this is a real option. But 1.7 bits per weight is already a different model. Before you use it, run it on your own tasks against the API of the original.

TOPIC 5

A license plate opens a car: GPT 5.6 helped the researchers

SOURCES Shubham Shah, 07.11 Shubs · HN, 08.11 Hacker News · The Guardian, 08.11 Guardian

Shubham Shah and Sam Curry hacked the connected-car platform of a large manufacturer. They do not name the brand. The attack needs only the license plate of the car.

The chain is long, but each link is simple. The plate gives the VIN, and the VIN gives the account of the owner. In the JavaScript bundle of the vendor, there was an admin login to the OpenIDM account system. A WAF bypass with path traversal gave access to its API. Then the researchers changed the password of the owner, bypassed MFA and got a token of the mobile app. This token opens the doors and starts the engine. Together with the second write-up by Curry, the attack worked in Australia, the USA and Canada.

A model found the key WAF bypass. People tried the prefixes by hand, got 403 and decided that everything was closed.

GPT 5.6 Sol Ultra read the OpenIDM code for six hours and found the prefix `/static%2f..%2fopenidm/`, which the WAF did not check.

The Guardian writes on the same day about the wider picture. Richard Horne, head of the UK NCSC, says that the world is only in the "foothills" of a new wave of attacks. According to the paper, large companies have found eight times more vulnerabilities since April. NCSC founder Ciaran Martin disagrees: this is a bad series, but not an unprecedented one.

Why it matters: a model does not get tired of trying variants that a person dropped as obviously closed. A WAF in front of an admin panel that is open to the internet now protects even worse than before. Find public admin panels and secrets in JS bundles before the model of someone else finds them.

TOPIC 6

Texts from AI clog the work of public bodies

SOURCES The Guardian, 08.11 Guardian · Semafor, 08.11 Semafor

The Guardian collected complaints from British public bodies. An employee in the office of an MP says that half of the letters they get are written with AI. A complaint to a local authority took about two pages before. Now most are 10 times longer. The largest known

complaint has 200 pages. Transport for London got a record 4,769 freedom of information requests in one year, 5% more than the previous peak.

The biggest problem is that AI writes a complaint like a lawsuit. The text cites laws and court cases, sometimes from another jurisdiction, and sometimes invented. But a person must still answer. Researchers from Hertie School, the Cooperative AI Foundation and GovAI already named this "agentic flooding".

Semafor showed the same mechanism in the press. A campaign against Waymo placed op-eds in local newspapers. The detector Pangram rated two op-eds in the Dallas Morning News as 95% and 71% written by AI, and an op-ed in the San Francisco Chronicle as 88%. None had a label. One of the authors admitted that a PR firm wrote the first version.

Why it matters: AI makes text cheap for the author, but not for the reader. Public bodies with a fixed budget feel this asymmetry first. The scores of the detector are a probability and not a proof, thus the labels of authorship stay the responsibility of the publication.

TOPIC 7

Science: EEG headsets feed brain models with user data

SOURCES Science, 05.11 Science · HN, 08.11 Hacker News

Science investigated consumer neurotechnology for one year. The Canadian company Interaxon, maker of the Muse headband, trained a "brain model" on about 50,000 hours of EEG data of its users. Its privacy policy did not say this directly.

The company updated the policy when Science pointed it out.

The journal checked the privacy policies of 57 companies that collect EEG. Four have no policy at all. Nine companies publicly admitted that they build a foundation model of the brain. Only six directly promise not to sell brain data. More than two thirds allow a transfer of the data to the buyer when the company is sold.

The regulation is still patchy. Chile protected brain data in its constitution in 2021, and in 2023 its Supreme Court ordered Emotiv to delete the data of a former senator. Colorado and California added neural data to sensitive personal data.

Why it matters: it is almost impossible to remove data from a model after it goes in. Thus the consent must be separate and come before the training, and it must not be a policy edit after an article by journalists. Companies already say directly how much this data is worth: "Our brain data is the most valuable part of the company", says the marketing director of Interaxon.

TOPIC 8

Outlook.com started to reject some mail from domains without DMARC

SOURCES MailRoute, 06.11 Mailroute · HN, 08.11 Hacker News [single source]

Since 05.11, the Microsoft mail services for private persons (outlook.com, hotmail.com, live.com, msn.com) reject some messages from domains that have no DMARC record. The sender gets `550 5.7.1 Message rejected: Message lacks SPF/DKIM/DMARC Auth Records`. MailRoute, a mail forwarding service, writes this from its own logs.

The danger is that the rejection is partial. One message arrives, the next one bounces, and a retry can pass. Before, Microsoft required SPF, DKIM and DMARC only from domains that send more than 5,000 messages per day. Now the rule also affects small offices.

MailRoute writes that Microsoft announced nothing. A commenter on HN disagrees and gives a link to the Microsoft documentation. The minimal fix is one TXT record `v=DMARC1; p=none`, but SPF or DKIM must pass for the domain of the sender.

Why it matters: a small business that sends invoices and confirmations to personal addresses loses messages in silence. Large mail services have moved toward DMARC as a minimum for everyone for a long time. You can check your domain with one `dig` command.

TOPIC 9

Collison: nobody has yet described a positive AI scenario

SOURCES @patrickc, 08.11 @patrickc · @gdb, 09.11 @gdb · @levie, 09.11 @levie · The Wall Street Journal, 08.11 WSJ

Patrick Collison, co-founder of Stripe, published two questions about AI. The post got more than 530 thousand views.

The first question: the risks of AI are well described, but the good scenario consists only of ingredients, for example "cure diseases". Do people work in such a world, what is scarce, and who is in control? Collison quotes Iain Banks:

the people in his "Culture" are "somewhere between passengers, pets and parasites".

The second question is about the moral weight of AI. If you give models even a tiny moral weight, then with quadrillions of instances their total weight outweighs humans in any utilitarian calculation. Collison writes that then one of three assumptions must go. Today people argue about data centres, and he asks how the support for AI changes when the talk reaches these questions.

The same evening, Greg Brockman wrote that frontier models with large compute at inference become "the engine of scientific discovery". Aaron Levie added that compute now limits productivity, and human time does not. The WSJ gives the opposite view from pharma: no drug made by AI has yet got FDA approval, and pharma executives say that biology is the brake.

Why it matters: the promises that "AI will cure diseases" sound more often, but people check them slowly. The question of Collison about what society looks like after this has no answer yet from any lab.

TOPIC 10

Agent memory: nine formats, and none reads the others

SOURCES LoreKit, 08.11 Lorekit · HN, 08.11 Hacker News

The author of LoreKit counted where coding tools store what they "remember". The places are AGENTS.md and CLAUDE.md, Cursor rules, the `/memories` folder of the Anthropic memory tool, Copilot Memory, Letta, Mem0, LangGraph, Graphiti graphs and MCP. Each format is smart on its own, but none reads the others.

The author compares this with monitoring before 2019. Then each vendor had its own format, and OpenTelemetry gave a shared data model. Vendors stopped competing on the format and now compete on analysis. For memory, a neutral place already exists: MCP and AGENTS.md belong to the Agentic AI Foundation at the Linux Foundation.

Useful details from the list. Copilot Memory deletes a fact that nobody used for 28 days, and it links facts to quotes from code. Graphiti stores two times: when the fact was true and when somebody recorded it. The author states honestly that he sells one of the formats himself.

Why it matters: a lesson that an agent learned in CI is not visible now to an agent in the editor. If you choose a tool today, check if you can export the memory from it.

---

misc

- **C++26 is stuck:** Herb Sutter writes that the DIS ballot failed for the first time in 36 years. The reason: Bjarne Stroustrup spoke against the contracts that the WG21 committee supported. Isocpp · HN Hacker News
- **KetchUp:** an open alternative to SketchUp, made with Rust and WebGPU, runs right in the browser. It has 292 points on HN.
  Github · HN Hacker News
- **Anthropic and gas:** the FT writes that Anthropic signed a $3.4 billion deal on gas power generation for its data centres. The text is behind a paywall, and only the headline is available. FT [single source]
- **Eval against an article:** @garrytan checked with his eval project an article about a 90% cost cut for agents with Kimi K3. The results of the article repeated. @garrytan

---