AI DIGEST

Investors and Musk demand the return of Huang to Nvidia

Also: Microsoft 365 outage, US flight stop, ChatGPT Wallet, Claude reads Gödel notes, llama.cpp 1.0, Edge0-145B.

Saturday, 14 November 2026

IN THIS ISSUE

1. A campaign for the return of Huang: an investor letter, a petition and Musk
2. Singapore checked Alice Huang for fraud and did not investigate espionage
3. Microsoft 365 was down for almost a day after a network capacity upgrade
4. The FAA stopped all departures in the US for the third time in history
5. ChatGPT Wallet: agents pay with a token, and the card stays in the bank
6. Claude helped to read 1,009 pages of Gödel notes: 663 theorems
7. llama.cpp released version 1.0.0
8. Edge0-145B: a model with 145 billion parameters on a laptop with 29 GB of memory
9. OpenAI disbanded its separate robotics team
10. Anthropic on safety evals: 10 rules and honest error intervals

TOPIC 1

A campaign for the return of Huang: an investor letter, a petition and Musk

SOURCES Stand with Jensen, 14.11 Standwithjensen · @altcap, 14.11 @altcap · @elonmusk, 13.11 @elonmusk · @balajis, 14.11 @balajis · @levelsio on Polymarket, 13.11 @levelsio · @amasad, 13.11 @amasad

Yesterday Huang left the post of Nvidia CEO, and WIRED wrote about an internal employee post in his defense. In one day, the defense moved outside the company. The site Stand with Jensen collects signatures under an open letter to the board. The letter demands the return of Huang until an independent investigation proves that his removal helps the company. On the morning of 14.11, the site shows 39,749 names. Among them are 2,842 current Nvidia employees, 416 former employees and 22,261 shareholders. The site counts these numbers itself, and nobody outside can check them.

The site runs on Replit, and Replit founder Amjad Masad shares it himself. Elon Musk, Shopify founder Tobi Lütke and Airbnb co-founder Joe Gebbia announced their support. The Altimeter fund of Brad Gerstner sent the board its own letter with "minimum standards" that the board owes to shareholders. On X, the letter text is an image, and we did not analyze it. By the words of Balaji Srinivasan, Altimeter and Atreides together manage

more than $25 billion of assets. The prediction market Polymarket gives 31% that Huang returns before the end of next year.

The argument of the supporters is simple: Nvidia without Huang will lose value. Some of them claim that the espionage story falls apart (see item 2). Skeptics in the replies write a different thing. The board probably would not remove the founder so suddenly without serious reasons. The campaign does not refute the facts about the affairs.

Why it matters: When a board removes a founder, this rarely becomes a public campaign with tens of thousands of signatures. The closest example is the return of Sam Altman to OpenAI in 2023. Then the employee threat to leave decided the case. Here the pressure comes from shareholders, and formally the board answers to them. Nvidia will publish its quarterly report on 17.11. This is the first public point where the board must talk to the market.

TOPIC 2

Singapore checked Alice Huang for fraud and did not investigate espionage

SOURCES BBC, 13.11 BBC · Singapore Ministry of Home Affairs, 31.10 Gov · Semafor, 13.11 Semafor · @balajis, 13.11 @balajis

On 13.11, the government of Singapore published a statement about the Megaspeed case, the BBC writes. Since February 2025, the authorities check if Megaspeed and other firms gave false declarations for servers that went through Singapore. By the documents, the equipment went to Malaysia, but investigators suspect other end recipients. Alice Huang led Megaspeed. The authorities checked her for possible participation in fraud. They did not investigate espionage. They found no evidence of a crime. The authorities kept her passport for several months and released her in February 2026.

Singapore also writes that it asked the US for help several times about Supermicro, which supplied servers to Megaspeed. By the words of the government, it got "limited information or assistance". There are still no charges in the Megaspeed case.

On X, Balaji Srinivasan shared another statement of the Singapore Ministry of Home Affairs as a new one. In fact, it has the date 31.10. In it, the government writes that it did not tell the WSJ about suspicions against Alice Huang.

It also writes that it did not investigate her links with Chinese intelligence. The document is real, but old. Its new wave on X is a part of the campaign from item 1.

Meanwhile, ministers in Washington argue. Treasury Secretary Scott Bessent told CNBC that his department asked for months to put Megaspeed on a blacklist. By his words, the decision belongs to the Commerce Department of Howard Lutnick (Semafor). Lutnick agreed to testify in Congress.

Why it matters: The Singapore statement does not close the question of chip smuggling. It removes only one part of the story: the claim that local authorities suspected Alice Huang of espionage. The case about server declarations stays open. Readers should keep three lines apart. The first is the affair and the CEO removal. The second is espionage, with weak evidence. The third is chip export, where the investigation continues.

TOPIC 3

Microsoft 365 was down for almost a day after a network capacity upgrade

SOURCES BBC, 13.11 BBC · Ars Technica, 13.11 Ars Technica · HN, 13.11 Hacker News

Yesterday we wrote about the outage of Outlook, Teams and Copilot. It continued for almost 24 hours. Microsoft admitted the problem on 12.11 at 17:54 GMT. It announced wide recovery only on 13.11 in the afternoon. The outage hit more than 70 countries and products that more than 430 million people use (BBC). Among the victims are the government of Japan, the Supreme Court of the Philippines, part of the NHS in Britain and the US House of Representatives.

Microsoft writes that the cause was a technical defect from a "network capacity upgrade". The company separately states that the outage has no link to a cyberattack or to AI agents. Ars describes the mechanism from status updates.

Servers failed automatic health checks and dropped out of rotation. Some returned servers froze again as soon as traffic came to them. Microsoft extended the mail delivery timeout to 48 hours, so that no mail is lost. The company promises a full report on the causes after the recovery ends. Ciaran Martin, former head of the British NCSC, said that the outage was possibly worse than CrowdStrike and AWS.

Why it matters: The mechanism here is typical for large clouds. Protective automation sees "sick" servers and removes them. The remaining servers get more load and also become "sick". Thus the recovery is slow and goes in parts. An organization that keeps all mail and chats with one vendor needs a backup communication channel for such a day.

TOPIC 4

The FAA stopped all departures in the US for the third time in history

SOURCES Ars Technica, 14.11 Ars Technica · Semafor, 14.11 Semafor · Ars Technica, 13.11 Ars Technica · HN, 14.11 Hacker News

On 13.11 at 15:36 Eastern time, the FAA stopped departures from all US airports. Transportation Secretary Sean Duffy named a "communication issue" in all air traffic control centers as the cause. The Department of Homeland Security said that there are no signs of an attack. The stop continued for more than 4 hours, and departures started again at 20:35.

By Semafor data, it hit more than 14,500 flights. This is the third such case. The first was on 11.09.2001, and the second was in January 2023, after a failure of the NOTAM system.

Duffy blamed "decades-old analog equipment". He promised that such stops will occur again without large upgrades. He asks Congress for $50 billion for modernization. Ars separately recalls the context. On 10.11, the FAA extended the AI system SMART to all 22 US centers. The system advises controllers on routes. No official links the stop to SMART. The controllers union said earlier that nobody included it in the development and tests of the system.

Why it matters: In two days, two different critical systems failed, and both failed without an attack. Microsoft broke during an upgrade, and the FAA failed because of old equipment. The practical conclusion: the most frequent cause of a large outage is your own changes and technical debt. Hackers are far behind. The request for $50 billion shows the price of debt in infrastructure that nobody updated for years.

TOPIC 5

ChatGPT Wallet: agents pay with a token, and the card stays in the bank

SOURCES OpenAI, 13.11 OpenAI · @gdb, 13.11 @gdb · Brave, 13.11 Brave · WSJ, 13.11 WSJ · HN, 13.11 Hacker News

Yesterday Replit let agents buy for business through Link and Visa. Today OpenAI launched ChatGPT Wallet together with American Express, Mastercard and Visa. The user adds a card one time. Before each purchase, the user sees the items, the sum and the card. Then the user confirms with Face ID or a fingerprint. The payment network issues a token in place of the real card number. Thus neither the agent nor the shop sees the number. Cashback and dispute protection stay the same as for a usual card. At first, Wallet is available in "dots" for ChatGPT Pro subscribers in the US, from age 18.

OpenAI writes directly that some sites will block such purchases, and anti-fraud systems will reject them as suspicious. Thus OpenAI also released a guide for business together with Wallet. The guide advises to let signed agent requests through (Web Bot Auth). It also advises to make agreements with Stripe, Adyen and anti-fraud providers.

Later, OpenAI wants to allow full delegation with limits on the sum and on categories.

On the same day, Brave opened its search API for payment by agents through the x402 protocol. A request without payment gets an HTTP 402 response with the price. The agent signs a payment in the USDC stablecoin on the Base network and sends the request again. The price is $0.005 per request, and the agent needs no account and no API key. The WSJ writes that banks cannot keep pace with agents. Stripe bets on agent commerce, and JPMorgan is skeptical. Amazon blocked purchases by the Muse agent from Meta.

Why it matters: In one week, two different paths for agent payments appeared. Card networks give a token and ask a person to confirm each purchase. The crypto path (x402)

removes the person and the account completely, but it works only for small API requests. For an online shop, one practical question exists now. Does its anti-fraud system block legitimate agents together with bots?

TOPIC 6

Claude helped to read 1,009 pages of Gödel notes: 663 theorems

SOURCES The Guardian, 13.11 Guardian · HN, 14.11 Hacker News

Since 2017, the Finnish logician Jan von Plato studies 16 working notebooks of Kurt Gödel from the early 1940s. They have 1,009 pages, and most of them use the old Gabelsberger shorthand. Von Plato says that six people in the world can read it, and he is one of them. The team read about one page per week. In September, von Plato gave Claude his transcriptions, six of his books about Gödel and related literature. In a few days, the model produced 650 pages of explanations.

The first task: do the notebooks contain a proof of the independence of the axiom of choice? In 1966, Gödel said that he proved it in the 1940s, but he could not reproduce the proof. The answer: almost. Gödel wrote one definition incorrectly, but he had the correct definition two years earlier. Von Plato checked the reconstructed proof in the Coq system, and it passed. Then Claude extracted all theorems with a proof or a sketch, and there are 663 of them. In Coq, about 400 passed as they are, about 200 needed corrections, and about 50 were false.

The logician Joel David Hamkins did not see the reconstruction. He says that the result is interesting if it survives a check. But historians must decide how much comes from Claude and how much from Gödel. "You don't get credit for work you don't tell anyone about," Hamkins adds.

Why it matters: This is a rare example where a model works on the full corpus. A person keeps a few dozen pages in mind, and the model sees a thousand pages together. An important detail of the setup: each conclusion goes through a formal check in Coq. The 50 false theorems show that a summary alone is not reliable without this check.

TOPIC 7

llama.cpp released version 1.0.0

SOURCES Hugging Face, 13.11 Hugging Face · @ggerganov, 13.11 @ggerganov · @ClementDelangue, 13.11 @ClementDelangue

3.5 years after the start, llama.cpp and the ggml library got version 1.0.0. The main change is a promise of stability. From now on, incompatible changes of the API and ABI increase the major version. New models, backends and features go into minor versions. The project supports more than 150 variants of model architectures and all common methods of

speculative decoding. NVIDIA, AMD, Qualcomm, Arm, IBM and Intel support the test infrastructure with their own runners.

The team writes separately about LLMs in its own development. They use models "thoughtfully", but control of the architecture and understanding of the implementation stay the priority. Together with the release, Hugging Face showed Llama desktop applications for macOS and Windows.

Why it matters: Many local AI applications rely on llama.cpp, from Ollama-like wrappers to mobile SDKs. Until now, any update could break the interface without a warning. Semantic versioning gives a simple rule to teams that put local models into a product. A major version means a check, and a minor version means a safe update.

TOPIC 8

Edge0-145B: a model with 145 billion parameters on a laptop with 29 GB of memory

SOURCES Edge0, 13.11 Edge0 · HN, 13.11 Hacker News

Edge0 released the open weights of Edge0-145B and Edge0-120B together with an inference engine. By company data, the 145-billion model runs on a laptop such as a MacBook Pro. Peak memory is 29 GB, generation is more than 80 tokens/s, prompt processing is more than 1,100 tokens/s, and the context is 128-256 thousand tokens. At each step, only about 10 billion parameters work.

There are four tricks. Each group of 64 experts shares two large matrices, and each expert has only small differences of its own. The team trained the model for local use from the start, without compression of a cloud model. The weights sit on three levels. "Hot" experts stay in GPU memory. A library of experts of ~40 GB stays in RAM. Another ~48 GB of dictionary data stays on the SSD. The CPU and GPU compute in parallel.

The company gives its own benchmarks: 64.6 on DeepSWE 1.1, 86.52 on Terminal-Bench 2.1, 90.4 on GPQA-Diamond. For comparison, it uses DeepSeek-V4-Flash, Qwen3.8-Flash-Next and Claude Opus 4.6. These are older cloud models. The company promises to publish a technical report with the test settings.

Why it matters: Until this autumn, most local models had fewer than 10 billion parameters because of memory limits. The split of weights between GPU, RAM and SSD removes this ceiling, and the cost is a more complex engine. If independent tests confirm the numbers, tasks with sensitive data will get a real local option. An example is work on code without network access. So far, all numbers come from the company.

TOPIC 9

OpenAI disbanded its separate robotics team

SOURCES Fortune, 13.11 Fortune · HN, 13.11 Hacker News

Fortune reported, and OpenAI confirmed, that the separate robotics team no longer exists. The researchers moved to the main research team of Mark Chen. The applied staff and the hardware moved to the Applied organization under Greg Brockman. The OpenAI booth at the CoRL conference in Austin was empty. The company says that it does not abandon robotics. Its explanation: general models, among them GPT-6 Astra, can control robots directly. In one Nvidia study, an agent on Astra had 74.8% success in manipulation simulation. The best specialized models had 51.0%.

Fortune adds context. OpenAI cuts "side quests" and earlier closed Sora. By reports, the loss in the last quarter was $10.9 billion. MIT professor Russ Tedrake disputes that labs win automatically. Robotics companies have data from real factories, and better base models make this data more valuable.

Why it matters: If a general model controls a robot without separate training, the billions in "robot brain"

startups stand on a weaker base. But there is a limit. A large model runs in the cloud and slowly, and a robot needs a fast reaction without a network. Thus the most probable combination is this: a general model plans, and a small local model executes.

TOPIC 10

Anthropic on safety evals: 10 rules and honest error intervals

SOURCES Anthropic, 13.11 Anthropic · @AnthropicAI, 13.11 @AnthropicAI

In its engineering blog, Anthropic published 10 recommendations on how to build, run and report safety tests of models. Some points with numbers follow. Single-turn tests are saturated. On the API, all recent models give 94-98% safe answers, so these tests no longer show a difference between models. Multi-turn tests give a wider spread, because harmful intent often shows only after several messages. An example is a test on romance scams, where no single message breaks the rules.

The minimum sample by Anthropic: 100 scenarios for a complex multi-turn test and 1,000 for a simple test. The error interval on multi-turn tests can be wider than ±10 percentage points. The company gives its own example. Claude Opus 5 and Opus 5.5 showed 98% and 94% on the scam test, but the difference is not statistically significant. About grader models: a team must compare them with human labels, and a broken grader must fail with an error. A silent score from a broken grader is worse than no score.

Why it matters: The recommendations apply to any system where one model grades another. A typical error is to compare two numbers without an interval and to announce an improvement. The example of Opus 5 against 5.5 shows that a difference of 4 points can be noise.

misc

- **Anthropic IPO:** Semafor writes that the order book is six times oversubscribed. Goldman Sachs, one of the lead banks, forecasts revenue of more than $600 billion in 2030. The price is expected on 18.11.

  Semafor
- **Dave Ricks against the forecast of Dario:** on 11.11, Dario Amodei said that cancer, Alzheimer's and mental illness will be in the past in less than 10 years. Eli Lilly CEO Dave Ricks answered that this is "not likely". Biology is still "almost a blank page", and clinical trials on people stay the longest stage. Gary Marcus collected the quote.

  Substack · HN Hacker News
- **Windows SignDrop:** a researcher with the alias Nightmare Eclipse published a 0-day in the Common Log File System.

  A usual user gets SYSTEM rights through Windows Defender. Will Dormann reproduced the exploit, but only on some Windows 11 26H2 systems. Microsoft says that it sees no attacks and that Defender blocks them. Ars Technica
- **TSMC:** in a report to the SEC, the company said that data of some current and former employees appeared on a third-party site. The data matches a system of a cloud vendor. TSMC found no signs of a breach of its own systems.

  Sec
- **Claude:** on 13.11, claude.ai, Claude Code and Console had a login outage. Anthropic fixed it and did not name the causes. Claude · HN Hacker News
- **Google DeepMind:** the lab opened the PARDIG and SatWeave datasets, which trained WeatherNext 3. PARDIG estimates precipitation over the full planet every 15 minutes. SatWeave combines geostationary satellites into one picture of clouds every hour. @GoogleDeepMind
- **@lennysan:** 15% of all open tech jobs are now in the San Francisco Bay Area, and this is a record. For product managers and designers, the share is even higher, one in four. @lennysan