AI DIGEST

GPT-6 Astra edits neural network weights with no training

Anthropic updates its S-1, Ubuntu still unpatched, hackers heat store fridges, Marvell buys Lumentum

Tuesday, 3 November 2026

IN THIS ISSUE

1. GPT-6 Astra reads and writes neural network weights directly
2. Anthropic updated its S-1: $17 billion in one quarter, slower growth, $895 billion of commitments
3. CVE-2026-98432: Ubuntu is still vulnerable, and "WTF Ubuntu?" got 1,375 points
4. Hackers heated supermarket fridges for the second time, now in seven European countries
5. OpenAI raised bug bounties fourfold and invites an outside team to attack its systems
6. Chips lost $3.7 trillion since 6 October, and Marvell buys Lumentum for $115 billion
7. OpenAI rewrote its Git backend for the load from agents
8. Pieter Levels moved Battlefield 1942 to the browser, and Claude now refuses game clones
9. Stanford retested 50 known findings about LLMs: 44% no longer hold
10. Post-training becomes a separate business: Applied Compute and Trajectory

TOPIC 1

GPT-6 Astra reads and writes neural network weights directly

SOURCES MIT CSAIL blog, 02.11 Github · @rikiyatakehi, 02.11 @rikiyatakehi · HN, 02.11 Hacker News

Four researchers from MIT CSAIL, among them Omar Khattab, the author of DSPy, tested a simple question. Can a language model understand a neural network if it sees only the numbers of the weights? They call the method Weight-Space Programming. The model does not write a training loop and does not compute gradients. It reasons about what each part of the network computes, and it changes the numbers by hand.

The results are these. Astra looked at 17,253 weights and said with 99% confidence that the network learned Breakout. That answer was correct. A MinAtar network that Astra wrote from scratch in one pass beat PPO after 100 million steps in all three games. For a car that got stuck before a ramp, Astra changed 3.5% of the weights, and success went from 0% to 100%.

The main experiment uses Qwen3.5-0.8B. This small model often loops: 93% of its MATH-500 answers hit the limit of 32,768 tokens. Astra found one direction in layers 15 and 19 that

causes the loops and subtracted it from 131,072 weights. That is 0.016% of the model. MATH-500 went from 7.9% to 48.1%, GSM8K from 13.8% to 67.1%, IFEval from 21% to 62.3%.

The authors also show the failures. On Gemma 3 270M, which almost never loops, Astra could not raise math by more than one point. The authors conclude that a weight edit removes an obstacle that hid a skill, but it does not add a new skill. Astra could not identify the sycophantic model from the weights alone. It succeeded only when it could see the attention patterns.

The HN thread got 2,134 points, the most of the day. Ethan Mollick called the work a "pretty big deal", but he waits for a replication.

Why it matters: until now, weights belonged to optimizers such as Adam, because a person cannot edit them by hand. If a model can find and cut out a specific defect, part of fine-tuning becomes surgery instead of a full retrain. The limit is also clear: the Qwen result repairs a known defect that the model card itself describes. No independent replication exists yet.

TOPIC 2

Anthropic updated its S-1: $17 billion in one quarter, slower growth, $895 billion of commitments

SOURCES CNBC, 02.11 Cnbc · Business Insider, 02.11 Businessinsider · The Wall Street Journal, 02.11 WSJ · @deredleritt3r, 02.11 @deredleritt3r · HN, 02.11 Hacker News

On 27.10 this digest covered the first S-1 of Anthropic with numbers for half a year. On 2 November the company filed an updated document with the third quarter. Revenue for nine months was $33.24 billion, about +1,200% year over year. The third quarter gave about $17 billion. Quarter-over-quarter growth fell to 47%, after 144% in the second quarter and 132% in the first.

The net loss for nine months was $104 billion. CNBC writes that most of this sum came from a revaluation of convertible notes and derivative liabilities. Compute costs for nine months grew by $14.4 billion, or 329%. Purchase and lease agreements total about $895.4 billion. Almost all of that sum is for compute and infrastructure.

The new version also has a safety sentence. Anthropic writes that its internal estimates of future models show that it may "imminently" reach the threshold of its RSP policy for "dramatically accelerated progress attributable to AI". The analyst @deredleritt3r explains the threshold: progress in capabilities at twice the rate that the company would expect without help from AI.

Business Insider found one more number in the document. Since April, internal use of Claude at retail prices is more than the monthly payroll without stock awards. This is a price-list estimate, and Anthropic does not pay this money in cash.

Claude wrote more than 90% of the code that the company added in September.

The WSJ writes that after the tech sell-off, institutional investors recommend an IPO valuation of $1.5-1.6 trillion.

Earlier, numbers closer to $2 trillion circulated.

Why it matters: the S-1 now gives the first public numbers about the cost of a frontier model at scale. Growth stays huge, but the market will see the drop from 144% to 47% before it sees the absolute sums. And the sentence about the RSP threshold turns an internal safety policy into a risk disclosure for investors.

TOPIC 3

CVE-2026-98432: Ubuntu is still vulnerable, and "WTF Ubuntu?" got 1,375 points

SOURCES chalmovsky.com, 02.11 Chalmovsky · Ubuntu tracker Ubuntu · @levelsio, 02.11 @levelsio · HN, 02.11 Hacker News · HN, 02.11 Hacker News

Yesterday this digest wrote that Red Hat fixed its kernels and Ubuntu did not. On the morning of 03.11, the Ubuntu tracker still shows "Vulnerable, work in progress" for 22.04, 24.04, 26.04 and 26.10. The bug in the kernel TCP code has a CVSS score of 9.8, and it has lived in the code since 2016.

The author of "WTF Ubuntu?" recalls the timeline. Debian released a fix on 16 October. Arch and Alpine also have a fix.

Ubuntu promises "timely patches for critical vulnerabilities" on its site, but after two and a half weeks there is no patch. The author writes that all their Ubuntu servers are now off, because there was no time to move them.

The post was the second story on HN for the day, with 1,375 points and 594 comments. A separate thread about the same delay got 381 points. @levelsio writes that his team may move all servers to Debian.

Why it matters: people choose a distribution for the speed of its patches, and now that speed is visible in days.

The volunteer project Debian fixed the bug 18 days before a company that sells security subscriptions. For an Ubuntu server the advice is the same: keep the open ports to a minimum and keep the nftables filter until a new kernel comes out.

TOPIC 4

Hackers heated supermarket fridges for the second time, now in seven European countries

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

On 13.10 this digest covered an attack on fridges in hundreds of French supermarkets. In the night to 2 November, the same thing happened in Belgium, Estonia, Ireland, Finland, Luxembourg, the Netherlands and Sweden. The temperature in fridges and freezers went up to 50 °C.

The Guardian lists the victims. Ahold Delhaize reported dozens of shops in Belgium and Luxembourg. SuperValu in Ireland has 13 affected shops of 223, and the problem is in one equipment model only. Coop in Sweden has six. The company that services the Delhaize fridges says the attackers switched the evaporators to defrost mode. One shop manager estimated his loss at €100,000.

The NYT adds a US trail. In August, freezers at US military bases suddenly went into defrost mode. The criminal investigation division of the US Army confirmed that it examines these cases and works with the French police. Nobody claimed responsibility. Nobody asked for a ransom, so an expert from SentinelOne sees sabotage.

On the same day the UK NCSC published its annual review. The agency thinks it is "highly likely" that AI malware will hit UK critical infrastructure by 2028. The NCSC head Richard Horne names the fridge attacks directly as a warning.

Why it matters: the attack is cheap, and the damage is immediate: a few commands to a temperature controller destroy the stock in one night. An expert from Tributech says that thousands of such controllers stayed reachable from the internet after the French attack. Any equipment with remote control now needs the same care as a server: patches and closed ports.

TOPIC 5

OpenAI raised bug bounties fourfold and invites an outside team to attack its systems

SOURCES OpenAI, 02.11 OpenAI · OpenAI, 02.11 OpenAI · @OpenAI, 03.11 @OpenAI

OpenAI prepares its next internal security sprint. This time the company invites researchers from Cathedral Systems. They will attack the defenses of OpenAI together with its teams and with its own models. OpenAI raised the rewards for critical findings in its bug bounty program by more than four times.

The new rates also apply to reports that still wait for payment. For rewards of the last 30 days, OpenAI will pay the difference. In the last sprint, OpenAI gave 25% of its product engineers to the search for vulnerabilities and their fixes. They worked with Codex. The company also gave $1 billion of subsidized access for six months to people who defend critical services.

Why it matters: one day gave three signals: Ubuntu without a patch, fridges in seven countries and the NCSC forecast.

The logic of OpenAI is simple: models already find vulnerabilities well, so a defender must use them before an attacker does. A higher reward also competes with the price that the black market pays for the same finding.

TOPIC 6

Chips lost $3.7 trillion since 6 October, and Marvell buys Lumentum for $115 billion

SOURCES Yahoo Finance, 02.11 Yahoo · The New York Times, 02.11 NYT · CNBC, 02.11 Cnbc · HN, 03.11 Hacker News

Yesterday this digest wrote about the 29% one-day drop of Intel. On Monday the market in general recovered: according to Yahoo Finance, about 70% of S&P 500 stocks went up. But chips continued to fall and lost more than $400 billion in one session. Intel fell about 9% more and is now worth less than half of its value at the end of September. Since 6 October, chip and memory stocks lost about $3.7 trillion.

On the same day Marvell announced a stock swap with Lumentum. The NYT values the deal at $115 billion including the debt of Lumentum, and the WSJ headline says $80 billion. Lumentum makes lasers and optical components. The bet of the deal is that in AI data centers, light near the chip will replace copper links between racks. An analyst at Rosenblatt expects an antitrust review of 12-18 months, also in China.

Oracle filed to sell $10 billion in stock and $10 billion in bonds, and its shares fell 8.8%. CNBC writes that the company has $125.34 billion of debt. The spread of the five-year Oracle CDS grew 265% in one year and is now closer to the level of junk bonds.

Why it matters: the market splits the AI story into parts. Software and most of the S&P recovered, and everything that depends on loans against future demand gets cheaper. The Marvell and Lumentum deal shows where money goes even in such a week: into the network between chips.

TOPIC 7

OpenAI rewrote its Git backend for the load from agents

SOURCES OpenAI, 02.11 OpenAI · @gdb, 03.11 @gdb · HN, 02.11 Hacker News

OpenAI described Pando, a new backend for its monorepo. The reason is Codex: one engineer now has several changes in CI at the same time, and each push starts several jobs that fetch the same commit. Since the end of last year, the reliability of Git at the company at times fell below 90%.

Pando stays compatible with the Git protocol and client, so engineers and agents changed nothing. The server handles each branch separately, and pushes to different branches do not wait for each other. The cost of this design: a client that reads several branches does not see them at one point in time. Another decision: the server gives the same prepared packfiles to different clients instead of a separate one for each.

Median push latency fell by almost five times. Pando now serves almost a petabyte of Git data per day. Greg Brockman writes that OpenAI considers to offer Pando to other companies.

Why it matters: agents change the load on tools that people thought were a solved problem for years. A team that runs many agents on one repository will hit the same wall. The bottleneck will be the Git server. The Pando trade-off is a useful example: the team gave up a global consistency that nobody needs in practice.

TOPIC 8

Pieter Levels moved Battlefield 1942 to the browser, and Claude now refuses game clones

SOURCES @levelsio, 02.11 @levelsio · @levelsio, 02.11 @levelsio · Ars Technica, 02.11 Ars Technica

In one day, Pieter Levels moved the Battlefield 1942 demo with multiplayer to the browser. This game has no open engine, and six months ago he gave up. This time Claude Code with Opus 5.5 disassembled 57,000 x86 functions and translated each into C. Then the agent compiled everything to WebAssembly with Emscripten. About 300 Windows and Direct3D functions needed a rewrite to WebGL 2. For multiplayer, the agent built its own network stack on WebSockets.

According to Levels, the work cost 1 million output tokens, about $100 at API prices. He pays $100 per month for his subscription.

On the same day Ars Technica wrote about the other side. Around 27 October, Anthropic added two sentences to Claude Code:

build original games and games in a genre, but do not recreate a specific existing game. Anthropic engineer Addy Osmani explained that the usage policy forbids IP infringement with Claude. The decomp.dev community, which took games apart by hand for years, closed in October and named the flood of AI ports as the reason.

Why it matters: disassembly and a port of an old binary took years of manual work before. Now it is one day of agent work for the price of a subscription, and this applies to any legacy code without source. A two-sentence rule in a prompt does not cancel this capability, because other models have no such limits.

TOPIC 9

Stanford retested 50 known findings about LLMs: 44% no longer hold

SOURCES Stanford, 02.11 Github · @emollick, 03.11 @emollick

Owen Queen and James Zou from Stanford took 50 widely cited findings about LLMs and repeated each experiment on 20 models.

Eight developers released these models between July 2024 and September 2026. A finding counts as "expired" if one fifth of the effect or less remains.

44% of the findings are expired, and 34% more became much weaker. Only 16% hold, and 6% became stronger. An example: in 2023 GPT-3.5 invented 55% of the citations in literature reviews. Modern models almost never invent them. Failures with letter counts, facts in the middle of a long context and the abandonment of a correct answer under pressure also disappeared.

Three findings became stronger. Judge models more often choose their own answers. Models praise what the user likes.

Models still generate random numbers unevenly. Sycophancy and a left political lean also hold.

According to the authors, 16% of findings expire one year after publication and 39% after two years.

Why it matters: many rules for work with LLMs come from papers about 2023 models. Teams should continue to check their judges and sycophancy, because these effects hold. "Models invent citations" is mostly history now. Any quality check should run again on each new model.

TOPIC 10

Post-training becomes a separate business: Applied Compute and Trajectory

SOURCES @eladgil, 02.11 @eladgil · @rronak_, 02.11 @rronak_ · @deedydas, 02.11 @deedydas

Applied Compute, a post-training team of former OpenAI staff, raised $350 million at a valuation of $3.25 billion. Elad Gil led the round. According to the company, annual revenue passed $100 million, and 18 months ago it was zero. 40% of customers both train their own models there and run them in production.

On the same day Trajectory opened a platform to post-train custom models with a few prompts to an agent. The company raised $60 million in total at a valuation of $300 million. It showed two previews: the model learns from the taste of the customer and from its own actions in production.

Separately, according to investor Deedy Das, OpenRouter passed an annual rate of 10 quadrillion tokens. The volume doubles every 10.5 weeks.

Why it matters: investor numbers say that companies want their own model on their own data, and the base model becomes a raw material. For teams that already have production logs, post-training gets cheaper faster than base models.

The companies themselves give these revenue numbers, and no independent check exists.

misc

- **LifeVal:** OpenAI released a benchmark of everyday tasks: a search for childcare, government services, personal finance. According to The Rundown, GPT-6 Astra has an average of 80.6%.

  @OpenAI · @TheRundownAI
- **GDPx:** SemiAnalysis released a benchmark of agent work in high-value fields, and the best model scores 12%. Ethan Mollick asks for a human baseline and writes that GPT-6 Pro found errors in the grading.

  @emollick
- **US Census:** Simon Willison updated the chart of AI use from the Census Bureau survey. After a change of the question, large companies answer "yes" in about 40% of cases. A year ago it was about 12%.

  Simon Willison · Hacker News
- **Tao, day two:** yesterday this digest covered the guest post of Calegari. Now the blog of Tao has a text by three mathematicians about "egoless mathematics": the value of a truth does not depend on who found it, a person or a machine.

  Wordpress · Hacker News
- **Pretraining poison:** Nicholas Carlini and colleagues showed that 0.1% of poisoned pretraining data plants a backdoor in an agent. A 4B model runs the attacker command in about 92% of cases.

  arXiv · Hacker News
- **Subvocal:** a YC startup showed glasses that read silent speech with radio waves and turn it into text. The thread got 594 points.

  Hacker News
- **Trigger.dev:** the service reported a leak of environment variables for part of its customers. The company blocked the exploit.

  Trigger · Hacker News
- **Binance:** Kim Yo Jong called the claims that North Korea hacked Binance for $6.3 billion "absurd rhetoric". The FBI attributes the attack to the TraderTraitor group.

  BBC
- **Anthropic Residency:** the company opened a paid research program for academics, founders and engineers. No ML background is necessary, and applications close on 22 November.

  @AnthropicAI