QOR /AI

14th November 2026 · The week of 7–13 November

A WEEKLY PAPER ON ARTIFICIAL INTELLIGENCE

The week

Weekly issue: the stories of 7th–13th November 2026

POLICY

Jensen Huang quits as Nvidia's chief executive. Nvidia announced his resignation on 12th November, four days after he admitted to an affair with Alice Huang, a Megaspeed executive whom American officials suspect of spying for China. The *New York Times* and Reuters have also reported affairs with two Nvidia employees, and Colette Kress, the finance chief, is interim chief executive. [techcrunch.com]

Huang's exit may owe more to Nvidia's lending than to his affairs, a report suggests. The *Wall Street Journal* says Nvidia offered cloud providers billions of dollars in guarantees so that they could buy its chips, and that Kress, then finance chief, halved one for OpenAI discussed at $250bn. CNN reports that the intelligence gathered so far does not prove that Alice Huang spied for Beijing. [the-decoder.com]

Congress calls Huang and the commerce secretary to account. Huang has agreed to testify to Congress behind closed doors before Thanksgiving, and Senator Tim Scott wants him and Howard Lutnick before the Senate Banking Committee. Two Democrats, Ron Wyden and Richard Neal, are investigating the administration's handling of Nvidia's sales of H200 chips to China. President Trump allowed those sales in December 2025, reportedly after being briefed on Huang's relationship. [techcrunch.com]

Six Chinese models were trained on chips banned from export to China, a researcher says. Will DePue, formerly of OpenAI, says traces in the published weights show that DeepSeek, Moonshot AI, MiniMax and Ant Group used Nvidia's Blackwell chips. (See Editorial.) [the-decoder.com]

Nvidia told a congressman in private that it could track its chips. Bill Foster, a co-sponsor of the Chip Security Act, told *MIT Technology Review* that Nvidia's representatives called the bill's location checks feasible. (See Editorial.) [technologyreview.com]

America approved a Chinese buyer of Nvidia chips that seems to exist only on paper, the *New York Times* reports. The Commerce Department cleared New Matrix, a 17-month-old firm, to handle H200 chips, but reporters found "a hollow gray shell" at its address. Its owners have ties to a Chinese private-equity firm that America has blacklisted since 2024. [theverge.com]

American data firms sell about $500m of training material to Chinese labs, ChinaTalk finds. The newsletter names at least nine sellers, among them Mercor, Surge AI and Turing, whose products include reinforcement learning[1] environments, often the same ones built for OpenAI or Anthropic. Export controls barely cover such data, and the authors want those for coding and biology brought under them. [the-decoder.com]

Australia will make safety rules binding on frontier[2] AI companies. Firms that want to operate there will have to report incidents, let outside experts evaluate their models and publish safety frameworks, under a new regulator that can impose "significant penalties". A review ordered after an OpenAI agent read non-public files on a Medicare statistics portal also asks the government to consider making a failure to prevent harm a crime. [medianama.com]

New York opens registration under its AI safety law. Developers with more than $500m in annual revenue, such as Anthropic and OpenAI, must register by 1st January 2027, when the RAISE Act takes effect. They must then publish safety frameworks and report critical incidents within 72 hours, or face fines of up to $1m for a first violation and $3m for later ones. [techcrunch.com]

Eight countries and the leading labs sign a safety pledge in Paris. OpenAI, Anthropic, Microsoft and Mistral promise to set up shutdown mechanisms for the models they control and to support independent audits, *Le Monde* reports. The "Paris Call for AI Safety and Security" sets no deadlines or sanctions, and Germany did not sign. [the-decoder.com]

The European Commission proposes rules for AI agents that sign contracts for consumers. Its Digital Fairness Act would let users decide which of an agent's purchases they must confirm. (See Editorial.) [medianama.com]

United Airlines will stop AI agents from booking its flights. It says agents given customers' logins can expose personal data and make purchases nobody intended. (See Editorial.) [techcrunch.com]

American public-sector unions win limits on AI at work. A union in Oakland secured retraining instead of AI-related lay-offs, and Pennsylvania's governor promised SEIU Local 668, another public-sector union, human oversight in a trial of ChatGPT Enterprise, *The Decoder* reports, citing the *Guardian*. Nearly 70% of American government agencies already use generative AI, according to a survey by Ernst & Young. [the-decoder.com]

Helsing unveils software that prepares military targets for weapons. The German defence firm says hundreds of AI agents in Clarity turn sensor feeds, reports and open-source intelligence into target recommendations, and that a named person must authorise every handoff to a weapon. It competes with Palantir's Maven. [the-decoder.com]

MONEY

Orders for Anthropic's shares exceed the supply six times, *Bloomberg* reports. The firm is offering 825m shares at $70–78 each, which would raise up to $64bn and value it at $1.4trn at the top of the range. Seven companies, among them Google and Nvidia with $10bn each, will separately buy $26bn of non-voting shares, and pricing is due on 18th November. [the-decoder.com]

Banks shut Chinese and Hong Kong investors out of Anthropic's listing, *Bloomberg* reports. Several have turned down orders from there, and some of the people involved cite America's rules on exports of defence goods. Anthropic, which declined to comment, already bars Chinese-controlled firms from its services. [the-decoder.com]

OpenAI's quarterly revenue passes $10bn, but its losses grow. Sales rose by about half in the third quarter and the adjusted operating loss widened to $10.9bn, the *Wall Street Journal* and *The Information* report. Anthropic made $17bn in the same quarter, with an operating loss of $567m. [the-decoder.com]

Nvidia's shares fall and the first investor lawsuit is filed. The stock lost 4.3% on 9th November, the first trading day after the board opened its inquiry, and about 8% more after hours on 11th November. A suit filed in California says Nvidia misled investors when its spokesman called Huang's interactions with Alice Huang "not unusual or unexpected". [techcrunch.com]

Anthropic buys $3.4bn of gas turbines, the *Financial Times* reports. The Caterpillar generators, first ordered by OnePWR, a supplier that could not finance them, can produce up to 1.7GW. Anthropic has mostly rented its infrastructure, and the queue for new gas turbines stretches into the early 2030s. [techcrunch.com]

CoreWeave asks its customers to pay up front. The firm, which rents out computing power for AI, expects about $9bn in prepayments and deposits over the next three months, and says it will not borrow on unsecured debt markets at current yields. Revenue rose by 168% in the third quarter, to $3.66bn, but it lost $709m. [techcrunch.com]

Google takes over Crusoe's unfinished data centre in Wyoming. Google, which reportedly worried about the project's cost and timetable, plans to invest $17bn in the campus. Crusoe, a developer recently valued at $30.9bn, had meant to scale the site to 10GW. [techcrunch.com]

One of China's largest AI data centres has 7% of the computing power of the world's largest, Epoch AI estimates. From satellite images, the research group puts Huawei's campus at Horinger, in Inner Mongolia, at the equivalent of about 81,000 of Nvidia's H100 chips, against SpaceX's Colossus 2 in Tennessee. It expects the gap to last until at least 2028. [epochai.substack.com]

LABS & PRODUCTS

Anthropic's Claude Fable 5.2 takes first place on an independent ranking. Artificial Analysis gives it 65 points on its Intelligence Index, five more than Google's Gemini 4 Carbon, at about $27 a task at the highest setting. Mythos 5.2, the same model with fewer safeguards in cyber-security and biology, goes only to vetted organisations, and Anthropic says its agents formalised the proof of Fermat's Last Theorem in Lean[3] in 30 hours. [the-decoder.com]

Fable 5.2 completes 41% of real freelance jobs well enough for a client to accept. Scale Labs and the Center for AI Safety, which run the Remote Labor Index, score it at 41.3%, against 2.5% for the best system a year ago. Anthropic's Opus 5.5 and

OpenAI's GPT-6 Astra manage about 21%, and human experts grade the 240 projects. [the-decoder.com]

Google releases two Gemini 4 models in three days. Gemini 4 Carbon, now open to all paying developers at $2 per million input tokens, scores 60 on Artificial Analysis's index, two points behind Fable 5.2 at the same setting. Gemini 4 Hydrogen, a cheaper model meant to replace Gemini 3.8 Flash, beats the larger Gemini 4 Argon by 51 points to 50, Google says. [the-decoder.com]

OpenAI finds that its own training choices make models' reasoning harder to watch. Less of the reasoning of its models, from GPT-5 to GPT-6 Astra, shows up in their chain of thought[4], and its safety team blames choices such as architecture, *The Decoder* reports. Britain's AI Security Institute finds that Astra solves maths problems that take a person about 31 minutes without writing out its reasoning. [the-decoder.com]

A free Chinese model comes close to OpenAI's restricted one at breaking into networks, Vals AI finds. On its CyberOffense benchmark GPT-6 Sol, part of OpenAI's restricted Daybreak programme, completed 80.8% of objectives and DeepSeek's openly released V4.1 Flash 72.8%, at about $3 a run against about $22. [the-decoder.com]

Tencent's chief AI scientist warns of AI-made bioweapons but argues against closing models off. Yao Shunyu, formerly of OpenAI, says a mix of open and closed models lets defenders use the technology too, the *South China Morning Post* reports. Tencent plans to publish the weights of its Hy4 model and says its biological capabilities fall below the high-risk threshold. [the-decoder.com]

Almost 95% of apps built with Grok were open to anyone, a security firm says. Claroc, which sells security tools for AI-built apps, found 35,578 of 37,571 live apps readable without a login, partly because the publishing screen preselected "Anyone with the link". They included a patient's laboratory results, plaintext passwords and confidential documents from SpaceXSI, Grok's maker. [the-decoder.com]

Hackers who broke into Fireworks AI took chat transcripts from Notion's users. The attacker reached a development system at Fireworks, which runs open models for developers, through its cloud provider, and Notion says about 1% of its customers were affected. It is the second breach at Fireworks in two months. [techcrunch.com]

RESEARCH

Researchers at Tsinghua University shrink AI agents to four bits with almost no loss. The full-size model coaches the compressed one through its mistakes, and the result comes within a point of the original on five agent benchmarks. (See Science.) [arxiv.org]

Agents err less when each remembered lesson links back to the run it came from. A team at KAIST, a South Korean university, cut the share of tasks that memory made its agent fail from 6.5% to 4.5%. (See Science.) [arxiv.org]

Researchers at StepFun and Adelaide University merge expert models into one in half the training steps. The student learns how much of each expert's weights to mix into its own as it trains. (See Science.) [arxiv.org]

[1] **reinforcement learning** training by trial and error, in which a model is rewarded for results [2] **frontier** the most advanced, most capable AI models [3] **Lean** a programming language in which proofs are checked by computer [4] **chain of thought** the step-by-step reasoning a model writes out before it gives its answer

EDITORIAL

Anthropic says it held its model back but lets nobody check

The company says it held back the skills that speed up AI research, tells investors they are its edge and lets nobody outside test either claim

On 9th November Anthropic released Claude Fable 5.2 and said it had deliberately kept the model, during training, from improving at the skills that could most directly speed up AI development. At optimising code for AI chips and improving model designs, the company says, Fable 5.2 is roughly as good as Opus 5.5, its everyday model. The next day Anthropic began the roadshow for its stock-market listing. In the video for investors Tom Brown, a co-founder, says the company's gains in efficiency "can accelerate even further as we approach recursive self-improvement", and that Claude will learn to find them "entirely by itself".

Both claims can be true, since one model can be held back while the company expects later ones to compound. But both rest on evidence that only Anthropic has seen. On 5th November METR, an independent evaluator that spent 11 weeks inside the company, estimated that Anthropic's progress had roughly doubled in pace and that most of the speed-up came from its own use of AI. Anthropic redacted a majority of the evidence METR found most informative. Sam McCandlish, another co-founder, said that publishing what drives the speed-up "risks handing out a roadmap for others to go faster", and promised the company's own report on the pace of progress within weeks.

Brown says Mythos, the company's restricted model, moved Claude onto AMD's chips this year in three days, writing or porting almost all the low-level code "with basically no human intervention". Such a move used to take dozens of engineers months. Staff using the company's best internal model reported to METR that they were about 12 times as productive as they would have been without AI. Nobody outside Anthropic has tested whether Fable 5.2 was in fact held back, or by how much.

The listing will not supply the missing check. The founders hold a class of shares with 50.1% of the votes, and the Long-Term Benefit Trust elects a majority of the board, so buyers of the shares on offer will have little say. They have still asked for six times as many shares as are on offer. Insulating a company from its shareholders can protect a mission. It also removes one of the few checks on what the company says about itself.

Governments are starting to fill the gap. From January New York will make large developers publish their safety rules and report serious incidents within three days, but the testing stays with the firms. Australia has gone further. Its proposed law would oblige frontier companies to let outside experts evaluate their models, under a regulator with the power to fine, though the details are still to come. In Washington the pressure is rising. Alex Bores, a New York assemblyman who helps shape the Democrats' AI agenda, has asked Congress to "ban RSI [recursive self-improvement], this year if possible". Scott Wiener, who as a state senator wrote bills to curb the risks of advanced AI, has just won a seat in the House.

Australia's test is the right one, and the new Congress should adopt it. Any lab that says its models are speeding up their own development, or that it has held such skills back, should give an evaluator chosen by a public body full access to the evidence. The evaluator should have the right to say in public whether the claim holds, even where the details stay secret. Anthropic could offer METR that now, before its shares are priced on 18th November. A company asking the public for $64bn should not ask it to take its word on the claim that matters most. ■

WHO GAINS Anthropic, whose claims of restraint and of accelerating progress both go untested as it raises up to $64bn, and its founders, who keep a majority of the votes after the listing.

WHAT'S ABSENT Any outside test of the skills Anthropic says it held back, and the evidence that METR saw but could not publish.

SOURCES

1. The Decoder: Claude Fable 5.2 and Mythos 5.2 cement Anthropic's AI lead just before its IPO — https://the-decoder.com/claude-fable-5-2-and-mythos-5-2-cement-anthropics-ai-lead-just-before-its-ipo/
2. Simon Willison: Anthropic's IPO roadshow video — https://simonwillison.net/2026/Nov/11/anthropic-ipo-roadshow/
3. The Verge: Why a new report on Anthropic automating AI development is alarming researchers — https://www.theverge.com/ai-artificial-intelligence/1029126/anthropic-metr-report-accelerating-ai-development
4. TechCrunch: Anthropic sets IPO price range of $70 to $78 per share — https://techcrunch.com/2026/11/10/anthropic-sets-ipo-price-range-of-70-to-78-per-share-to-reach-an-implied-valuation-of-1-404-trillion/
5. The Decoder: Australia's proposed AI safety bill would force frontier model makers to open up to independent testing — https://the-decoder.com/australias-proposed-ai-safety-bill-would-force-frontier-model-makers-to-open-up-to-independent-testing/

EDITORIAL

Make Nvidia's chips report their location

Huang's fall is about one man, but its lesson is that export controls leak, and Nvidia has said in private that it can track its chips

Jensen Huang's departure from Nvidia began as a spy story and ended, for now, as one about governance. Nvidia said only that the board had accepted his resignation, and the reports on why point in several directions, from his affairs to the company's lending. Whatever the reason, the question that outlasts him is what happens to Nvidia's chips after they are sold.

This week brought the first technical evidence of where some of them went. Will DePue, a former OpenAI researcher who runs Unemployment Labs, a non-profit, says six Chinese models released since April carry the signature of Nvidia's Blackwell chips, which may not be exported to China. The trace comes from the way a chip fills a new model with random numbers before training, which depends on how many processing cores it has. His method identified all 24 control models whose hardware was known, and he says researchers at OpenAI and Anthropic have checked it. By his estimate Moonshot's Kimi K3 alone needed about 4,500 Blackwell chips for some 70 days. The evidence shows access, not smuggling, because Chinese labs can rent chips in data centres outside China.

Nvidia told *MIT Technology Review* in October that the diversion governments and the press had found was "a drop in the bucket" compared with China's own computing power. In March the Semiconductor Industry Association, of which it is a member, called the location checks proposed in the Chip Security Act "new, untested, and potentially infeasible". Yet Nvidia's Fleet Intelligence software can tell which country a chip is in, and its representatives told Bill Foster, a congressman backing the bill, that the checks could be done. "They never put anything in writing," he says.

Some argue that tighter rules would do harm. SemiAnalysis, a research firm, says that letting Chinese labs rent American chips abroad keeps the profits in American firms, lets America watch what the labs do and lets it cut them off at will. That argument needs tracking too, since nobody can watch chips whose location is unknown.

The Senate should put the location checks into its National Defense Authorization Act, choosing the version in which each chip reports its location when it comes online, which the Center for Security and Emerging Technology judged the most cost-effective. Nvidia should put in writing what it told Foster. Congress cannot regulate a founder's private life, but it can regulate where his chips go. ■

WHO GAINS Lawmakers pressing the chip-security bills, who now hold Nvidia's private word, and data centres abroad that rent Blackwell chips to Chinese labs while nobody checks.

WHAT'S ABSENT Nvidia's own account of where its chips are, and any response from the Chinese labs whose models DePue examined.

SOURCES

1. The Decoder: Former OpenAI researcher says he found Nvidia Blackwell fingerprints hidden inside six Chinese AI models — https://the-decoder.com/former-openai-researcher-says-he-found-nvidia-blackwell-fingerprints-hidden-inside-six-chinese-ai-models/
2. MIT Technology Review: Nvidia privately confirmed it could track chips to fight smuggling — https://www.technologyreview.com/2026/11/12/1148483/nvidia-chip-tracking/
3. TechCrunch: Export-banned Nvidia chips are behind China's most powerful AI models — https://techcrunch.com/2026/11/13/export-banned-nvidia-chips-are-behind-chinas-most-powerful-ai-models-uh-oh-jensen/
4. SemiAnalysis: Nvidia and the World Without Jensen — https://newsletter.semianalysis.com/p/nvidia-and-the-world-without-jensen
5. The Decoder: Jensen Huang's ouster likely traces back to Nvidia's risky business practices — https://the-decoder.com/jensen-huangs-ouster-likely-traces-back-to-nvidias-risky-business-practices-more-than-his-personal-life/

EDITORIAL

Shopping agents must say who pays them

OpenAI's new wallet lets ChatGPT buy for its users, and the European Commission's rule, that an agent with a paid interest must always ask first, should apply everywhere

On 13th November OpenAI launched ChatGPT Wallet, which lets its agents pay for purchases online. For now it is open only to adult Pro subscribers in America, and the agent asks each time. It shows the items, the total and the card, and the user approves with a fingerprint or Face ID. Card networks issue a stand-in number, so neither the agent nor the shop sees the real card. OpenAI plans eventually to let users hand purchases over entirely, within limits they set on spending, categories and time.

Three days earlier OpenAI opened ChatGPT, which it says 1.2bn people use each week, to businesses. Shops can register, add product feeds and set up Business Agents that answer customers' questions and help compare products, and these agents can also serve as targets for ChatGPT's advertising. An assistant that chooses on the buyer's behalf is thus being sold to sellers as well. Once it buys without asking, the buyer cannot tell whether a choice reflected the merits or a payment.

Sellers have worries of their own. United Airlines said this week that it will restrict unauthorised AI agents, because sharing logins with them "can expose sensitive information" and their actions "may result in unintended purchases, account changes, or use of MileagePlus miles". Bank of America has urged customers not to give AI assistants their passwords. Left alone, shopping by agent is likely to become a patchwork of blocked sites and private arrangements.

The European Commission has proposed a better answer in its Digital Fairness Act. On first use, a consumer would choose whether the agent must confirm every contract, only those that meet criteria such as a maximum price, or none. If the system has any payment or conflict of interest, it must disclose it and ask for confirmation whatever the setting. Shops could not oblige anyone to use an agent, and the maximum fine would be at least 4% of a trader's annual turnover in the countries concerned.

That conflict rule is the one that matters, and OpenAI and its rivals, among them Anthropic, Meta and xAI, whose agents already pay for things, need not wait for Brussels before adopting it. Any purchase in which the agent's maker is paid by the seller, through advertising or commission, should go back to the buyer with the payment shown. An agent that spends a person's money should answer to that person first. ■

WHO GAINS OpenAI and the card networks, which now sit between buyer and seller, and the shops that pay to be found by agents.

WHAT'S ABSENT Who pays when an agent buys the wrong thing, beyond the banks' usual protections, and any figure for how many people shop through agents.

SOURCES

1. The Decoder: OpenAI rebuilt AI shopping from scratch with ChatGPT Wallet — https://the-decoder.com/openai-rebuilt-ai-shopping-from-scratch-with-chatgpt-wallet/

2. The Decoder: OpenAI turns ChatGPT into a business directory with AI agents that handle customer questions — https://the-decoder.com/openai-turns-chatgpt-into-a-business-directory-with-ai-agents-that-handle-customer-questions/
3. TechCrunch: United Airlines blocks AI agents from booking flights, citing cybersecurity risks — https://techcrunch.com/2026/11/12/united-airlines-blocks-ai-agents-from-booking-flights-citing-cybersecurity-risks/
4. MediaNama: EU proposes Digital Fairness Act to protect online consumers — https://www.medianama.com/2026/11/223-eu-digital-fairness-act-explained/

LAB REPORTS

What the big labs announced. Every figure is the lab's own.

ANTHROPIC

Claude moves into the office

Anthropic has added Claude Sheets, for spreadsheets, Claude Whiteboards and Claude Data, which builds dashboards from tools such as Snowflake and Salesforce, to its app. Partners including Adobe, Atlassian and Salesforce can embed their apps in Claude. Claude Meetings, in beta for Team and Enterprise customers, joins Google Meet calls as a visible participant, takes notes and answers questions, and by default stays out of calls with outside participants. [the-decoder.com]

ANTHROPIC

A Claude plugin for factories

Claude for Manufacturing, a plugin for Team and Enterprise customers, connects Claude to Autodesk Fusion, Oracle NetSuite, Microsoft 365 and Infor. Anthropic says its 17 skills let Claude draft quotes from actual inventory, compare suppliers, write quality reports and, after human approval, order from backup suppliers when deliveries are late. It does not yet control machines, which Anthropic is testing separately in a research preview. [the-decoder.com]

GOOGLE DEEPMIND

DeepMind's agents check Signal's code

Google DeepMind says agents running on Gemini Carbon proved security properties of the compiled code behind the Signal protocol, which protects messages in Signal and Google Messages, writing more than 6m lines of Lean[3] in four weeks. One such property is that messages stay safe before and after a device is briefly compromised. Human experts set the goals, and the agents flagged about 1% of the work for human review. DeepMind says no widely used protocol's implementation had been formally verified before, by any method, and that its system caught two previously fixed bugs it put back into a copy of the code. [deepmind.google]

GOOGLE

Gemini picks from a list

Google's Gemini Decisions API scores a fixed set of options, such as which tool to call, and returns probabilities instead of text, in 150–250 milliseconds. It charges ten cents for each million tokens of input and nothing for output. Google says Gemini Decision 1, the model behind it, led its own benchmark suite with the lowest calibration error and was nearly five times cheaper than Gemini 3.5 Flash-Lite on it. [blog.google]

GOOGLE DEEPMIND

A humanoid sorts parts for Mercedes

Google DeepMind says Gemini Robotics 2.5, its model for controlling robots, does 30–100% better than its earlier versions, depending on the task. In a pilot on Mercedes-Benz production lines in Hungary and Germany, Apptronik's Apollo 2, a humanoid robot, sorted metal parts with a success rate of 96%, it says. DeepMind uses Gemini 4 Argon to generate simulated practice runs from real camera scenes. [the-decoder.com]

OPENAI

OpenAI sells attack and defence

OpenAI has launched two security products under its Daybreak brand, both running on GPT-6 Cyber. Daybreak Red Team, in early access, uses agents to probe web applications for vulnerabilities and to check reported flaws and patches. Daybreak Defense Factory scans a company's code for flaws and proposes fixes, with tests, for developers to approve. [the-decoder.com]

META

Meta's small model for agents

Meta has released Muse Ember, its smallest model, for the routine work of agents, at ten cents per million input tokens and 50 cents per million output tokens. Meta says it gives more intelligence per dollar than Claude Haiku 5.5 or GPT-6 Luna. On Artificial Analysis's test it scores 36, against 37 for Meta's Muse Spark 1.3, at about 14 cents a task against about $2.4, but Haiku beats it at real coding tasks. [the-decoder.com]

DEEPSEEK

DeepSeek opens its app's tools to developers

DeepSeek says the features of its app, among them web search with citations, search during reasoning, reading files and links, and safety checks, now work through its API on V4.1-Flash. Text recognition in images, speech recognition and speech synthesis are offered separately. Public-facing apps must switch on the safety checks, and developers stay responsible for moderation, model registration and labelling AI-generated content. [api-docs.deepseek.com]

ALIBABA

Qwen4 arrives in preview

Alibaba Cloud has announced Qwen4 and is offering a preview of Qwen4-Max through its Token Plan subscription and its QwenWork, Qoder and QwenStudio services and the Qwen app, *ITmedia* reports. It says the model is better at knowledge, coding and carrying out tasks on its own, but has published few details. [itmedia.co.jp]

PREFERRED NETWORKS

A small Japanese model for finance

Preferred Networks, a Japanese AI company, has published the weights of PLaMo 3.0 Fin, a small model for finance, under its PLaMo community licence. It is built on an 8bn-parameter model from joint research with Japan's National Institute of Information and Communications Technology and further trained on Japanese financial data. The company says such small models suit bulk work such as sorting customer questions, extracting information from documents and monitoring transactions. [preferred.jp]

[3] **Lean** a programming language in which proofs are checked by computer

SCIENCE & TECHNOLOGY

Compressed agents learn from their coach

A full-size model that coaches a four-bit copy through its mistakes brings the compressed agent within a point of the original on five agent benchmarks

Compressing an AI agent need not cost it its competence, provided the full-size model coaches it through its mistakes. Researchers at Tsinghua University's Institute for AI Industry Research and the University of Science and Technology of China call their method SteerDistill. On five benchmarks of agent work, from calling software tools to shopping on a simulated website, a version of Qwen3-8B with its weights stored in four bits averaged 53.2. The original model scored 53.5, and the best rival method for compressed models, QeRL, 51.0.

Storing each weight in fewer bits makes a model cheaper to run, which matters for agents because they call the model again at every step. But compressed agents fail in characteristic ways, the authors find. They overlook a requirement of the task or the state of their surroundings, for instance trying to read a file before entering the folder that holds it. Left to retry on its own, the compressed model found a correct action within four attempts at only 9.6% of such turns in one diagnostic test.

SteerDistill rewinds the task to the moment before the error. The full-size model, which has solved the same task, compares the two runs and tells the student what it overlooked and when the advice applies. The student tries again, and the step is kept only if it changes the environment in the same way as the teacher's run did. If not, the teacher refines its advice. With such coaching the student found a correct action at 82.3% of those turns. It is then trained to make the corrected choices without hints, so in use it runs alone.

The gains held across models. On Qwen3.5-4B and Google's Gemma 4 12B the method beat the best rival by about 1.7 points and stayed within a point of the full-size model. At two bits, where

every method lost much more, its lead grew to 3.4 points. It was also cheaper to train. On τ2-Bench, a test of customer-service agents, it beat QeRL by 1.8 percentage points with 67% fewer hours of computing, and its successful runs were shorter.

The method has requirements. It needs a run in which the teacher succeeded, an environment that can be restored to an earlier state and a reliable check of each step. The authors have not tested it where those are missing. A step that passes the local check does not guarantee that the task will succeed. If the result holds more widely, agents could run on cheaper hardware without becoming more careless. ■

SteerDistill: Quantizing LLM Agents via Closed-Loop Teacher Steering

Genming Cui et al. · Institute for AI Industry Research (AIR), Tsinghua University; University of Science and Technology of China
Posted 10th November 2026

Our full summary ai.org.ge/en/papers/2026-11-14-steerdistill

arXiv abstract · Full paper (PDF) · Hugging Face · Code

Agents that check their memories

Linking each remembered lesson to the run it came from lets an AI agent catch bad lessons, raising scores and cutting the tasks that memory makes it fail

An AI agent that keeps lessons from past tasks can fail tasks it would have solved without them, often by following a lesson that is wrong. Researchers at KAIST, a South Korean university, and DeepAuto.ai found that 52% of such failures traced to an item in memory that the agent had followed. Their remedy, WikiTrace, links each lesson to a summary of the run it came from and to the full record of that run. When a lesson lacks detail, conflicts with another or comes from a failed run, the agent can check the source before acting.

The researchers tested an agent built on DeepSeek-V4-Flash on eight suites from five benchmarks, including household tasks, science experiments and customer service. With WikiTrace it averaged 90.2, against 79.5 with no memory, and 3.2 points more than with the same lessons alone. The share of tasks it lost because of memory fell from 6.5% to 4.5%, a figure that average scores hide and the paper reports separately. The lessons are still inferred from runs marked only as success or failure, so a link lets the agent check a lesson without proving it right. And every lesson goes into the prompt, which grows with the memory. ■

Don't Just Trust Your Memory: Evidence-Linked Memory Reduces Memory-Induced Degradation in LLM Agents

Yukyeong Lee et al. · KAIST; DeepAuto.ai
Posted 10th November 2026

Our full summary ai.org.ge/en/papers/2026-11-14-wikitrace

arXiv abstract · Full paper (PDF) · Hugging Face · Code

Merging experts by learning the mix

Letting a student model learn how much of each expert's weights to blend into its own matches standard multi-expert distillation in half the training steps

Frontier labs increasingly train a separate expert model for each field and then fold them into one generalist by distillation[5], with each expert grading the student's own answers. The step is slow and tends to leave one field behind. Researchers at Adelaide University and StepFun, a Chinese AI company, with colleagues at three other universities, keep the experts' weights in play. The student is a mix of its own trainable weights and the frozen experts' weights, and it learns the proportions of the mix as it trains.

On Qwen3 models of 1.7bn and 4bn parameters, with experts in medicine, law, finance and following instructions, the method passed the standard approach's final score in half the training steps. At the end the smaller model closed 94.7% of the gap between the base model and the experts, against 80.9%. The weakest field rose from below half of the gap to nearly 90%. Each step costs up to a quarter more, but the larger model reached the standard method's final score in under a third of the time. The finished model runs no slower. Only small models and four fields were tested, and one variant fell about 20 points short at the larger size. ■

P^2-MOPD: Accelerating Multi-Teacher On-Policy Distillation via Parameter Priors

Sinuo Wang et al. · Australian Institute for Machine Learning, Adelaide University; StepFun (Shanghai); Shanghai Jiao Tong University; Singapore University of Technology and Design; Mohamed bin Zayed University of Artificial Intelligence
Posted 5th November 2026

Our full summary ai.org.ge/en/papers/2026-11-14-p2-mopd

arXiv abstract · Full paper (PDF) · Hugging Face · Code

[5] **distillation** training a smaller model to imitate a larger one

OFF THE RADAR

Work that matters, from writers the big outlets missed. Every claim is the author's.

Chinese experts set modest goals for AI talks with America

Xiao Qian, Lu Chuanying and three other Chinese experts, scholars at Tsinghua, Tongji and Fudan universities and a retired army colonel, interviewed by Concordia AI, a Beijing-based AI-safety group
AI Safety in China (Concordia AI) · *"Five Chinese Experts on How China and the US Should Talk About AI"* · Nov 13

Ahead of the next China–US talks on AI, the five say the most to hope for at this stage is a regular, predictable channel between the two governments. Xiao suggests agreeing which incidents to report and at what threshold, and Lu says reports should be handled by professional teams under strict confidentiality. They say China sees safety and development as reinforcing each other and expects to help write global rules on equal terms. [aisafetychina.substack.com]

WHY IT WAS MISSED published in a specialist newsletter on Chinese AI policy

A Brazilian editor reads Anthropic's prospectus for signs of a bubble

Sérgio Spagnuolo, co-founder and director of Núcleo Jornalismo, a Brazilian news outlet, and a former Knight fellow at Stanford
Núcleo Jornalismo · *"O teste econômico da IA chegou pra valer"* · in Portuguese · Nov 11

Spagnuolo argues that the listings of Anthropic and OpenAI will show whether AI is a bubble, because public companies must publish quarterly results. From Anthropic's filing he counts operating losses of more than $20bn from 2023 to mid-2026, computing costs that fell from 1,000% of revenue to 55% and at least $431bn of computing it cannot cancel. Five customers, he notes, provided 25% of its revenue in 2026. [nucleo.jor.br]

WHY IT WAS MISSED published only in Portuguese

WORTH READING

Patel brings his forecast forward

Dwarkesh Patel · Dwarkesh Podcast · *"The avalanche that hit math is coming for AI development"* · Nov 13

Patel explains why watching AI sweep through mathematics made him expect AI to match the best human AI researchers within a couple of years. [dwarkesh.com]

Marcus says the bubble is popping

Gary Marcus · Marcus on AI · *"The great Generative AI unraveling"* · Nov 13

Marcus reads the falls in Nvidia, Oracle and CoreWeave shares, and Huang's exit, as the start of the reckoning he has long predicted. [garymarcus.substack.com]

What Nvidia loses with its founder

SemiAnalysis · SemiAnalysis · *"Nvidia and the World Without Jensen"* · Nov 12

The research firm argues that Huang personally held together Nvidia's supply deals, its credit guarantees and its support for open models. [newsletter.semianalysis.com]

Almost general, still jagged

Ethan Mollick · One Useful Thing · *"Fable 5.2 and Jagged AGI"* · Nov 9

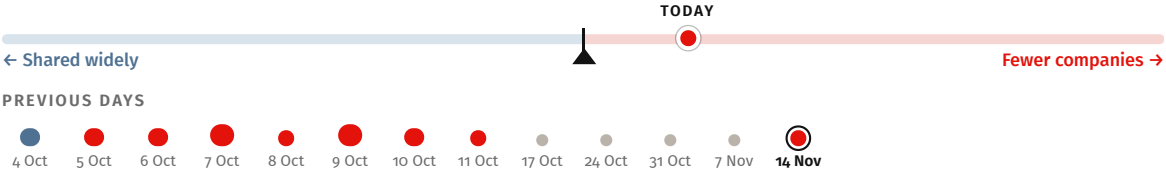
Mollick sets Fable 5.2 hard tasks, from hacking a keyboard to launching a business, to show where it matches experts and where it falls short. [oneusefulthing.org]

THE SCALES

Is power over AI held by fewer companies, or shared more widely?

Each day we look at the news and ask which way it moves this balance.

Slightly toward fewer companies



Last 7 issues: 3 toward fewer companies, 4 no clear change

Compared with 7 Nov: **moved further toward fewer companies**

Power over AI moved slightly towards fewer companies this week. Anthropic is raising up to $64bn while its founders keep most of the votes, bought its own power supply and asks the public to trust claims only it can check. Against that, Nvidia lost the founder who held its supply deals together, Chinese labs trained on chips meant to be kept from them, and Australia and New York moved to bind the labs by law. Of 49 stories, 32 are American, 10 Chinese, 3 European, 1 Japanese, 1 Korean and 2 from elsewhere.

HOW WE ASSESS THIS This is our editorial assessment, not a precise measurement. We read news from the US, Europe, China, Japan and Korea, and judge whether each story gives more control to a few companies or spreads it out.