NEW HORIZON · AI DIGEST

The Long Read

21 November 2026

12	3	4420	4
STORIES	SECTIONS	WORDS	HIGH IMPACT



4 high impact · 8 medium impact · spoke length = depth of coverage

The full-length companion to the daily New Horizon AI Digest.
Every story in today's email, reported at length · new-horizon.tech/digest

The issue at a glance

12 stories · 4420 words · 3 sections · 2 charted





AI Models & Research

01 Claude helps claim a Riemann proof as a human rival publishes

02 Inkling 2 Mini handles images and text on local hardware

03 METR finds AI coding gains do not guarantee good judgment

04 AI bug-hunting review finds more flaws, but no faster fixes

AI Tools & Ecosystem

05 Hugging Face cuts coding-agent token use with custom training

06 Databricks index shows identical coding scores can hide huge costs

07 Microsoft and 40 organizations back verifiable AI-agent actions

08 Copilotforce brings Salesforce customer records into Microsoft Copilot

AI Applications & Industry

09 Two million bait secrets expose leaks across AI services

10 Lenders resist financing SpaceX chips that cannot be repossessed

11 Spotify’s Daniel Ek backs a simulation-first fusion startup

12 Google commits over $500 million to super-hot geothermal power

How to read this. Every story in today's email is reported here at full length, in the same order. Impact is the writer's judgement of whether a story changes what a reader — in business, at work, as an investor or as a builder — should do or believe this week. Charts appear only where the source itself puts comparable numbers side by side; nothing is estimated to fill a gap. Sources are listed in full on the final page.

SECTION 1 OF 3

AI Models & Research

4 stories · 1 high · 3 medium

01 MEDIUM IMPACT TechCrunch

Claude helps claim a Riemann proof as a human rival publishes

> An Anthropic mathematician has announced a machine-checkable proof of the Riemann hypothesis, the 167-year-old problem about the hidden structure behind prime numbers, and said the AI model Claude helped him find it.

Levent Alpöge, a mathematician employed by Anthropic, published an essay saying he had found a Lean-compilable solution to the Riemann hypothesis, first posed in 1859 and one of the Millennium Prize problems. Lean is an open-source coding language that expresses mathematical arguments as code so a computer can check them step by step. Several participants in a Zulip chat on the result, including MIT mathematician Andrew Sutherland, said they had confirmed the solution compiles; the story was later updated to say independent sources had checked the Lean proof. Alpöge wrote that he messaged Anthropic chief executive Dario Amodei last Sunday night to ask how to announce it, adding: "The precise timing is horrible." The timing: the same week his employer became the first publicly traded frontier AI lab.

What the proof actually says is not yet clear to anyone. In the Zulip discussion, Chris Birkbeck, a number theorist at the University of East Anglia, described the argument as drawing on automorphy lifting, Galois representations, base change and a list of other machinery. Alpöge acknowledged releasing the result without a human-intelligible explanation, saying his obsession with the problem since childhood meant he "could not live" if he waited. He described working through it with Claude as a "stream-of-consciousness jaunt" through "incredible confusion," and said the model "flew, and danced, and sparkled near that finish line." He has committed to pursuing an explanation humans can follow.

The same week, Ravi Vakil, a Stanford professor and president of the American Mathematics Association, posted a proof of the General Riemann hypothesis written by Stanford postdoc Alessandro Masullo, noting that "it is all human, with no AI." Vakil said the plan had been to bring Masullo to Stanford for a year to polish the writing and circulate it before announcing; "now it seems best to put out the current draft." Masullo said the work took 12 years. Because it is not written in Lean, it is too early to say whether it checks out.

Alpöge and NYU mathematician Tristan Buckmaster also released another important proof, on the Navier-Stokes equations. For readers outside mathematics, the practical question is what a verified proof is worth before anyone can explain it: the checking is real, the understanding is not yet there.

KEY FACTS

PROBLEM AGE
Dates to 1859

PRIZE STATUS
One of the Millennium Prize math problems

PROOF LANGUAGE
Lean, an open-source coding language for formal verification

HUMAN RIVAL'S TIME ON PROOF
12 years

ANNOUNCEMENT TIMING
Same week Anthropic became the first publicly traded frontier AI lab

WHY IT MATTERS

Formal verification tools like Lean give AI-assisted mathematics a way to prove a result is correct even when no human can yet follow the argument — a check on AI claims that businesses relying on automated reasoning will increasingly want. The simultaneous human proof shows the race is now about credit and comprehension as much as discovery.

Read the original at TechCrunch →

02 MEDIUM IMPACT huggingface.co

Inkling 2 Mini handles images and text on local hardware

> Thinking Machines Lab has released Inkling 2 Mini, a compact open model that handles images and text and is built to run on a user's own machine rather than in a data centre.

The model is published on Hugging Face with support for the main open-source tools developers use to run models locally, including llama.cpp and MLX, the framework for Apple silicon. A four-bit version — compressed enough to fit on ordinary hardware at some cost in precision — is available for both the standard model and a research variant. SGLang, a serving system for data-centre deployments, supports it from day one.

The pitch is capability per gigabyte. Inkling 2 Mini is a mixture-of-experts model: rather than running every part of the network for every word, it routes each token to a small subset of specialist sub-networks. The source says it has about 26 billion parameters in total but activates roughly 3 billion per token, which is the figure that governs how much compute each answer costs. It also reuses a block of middle layers three times in a loop, so a 24-layer model behaves like a 48-layer one without storing twice the weights. In plain terms: a model that punches above its weight class while fitting on a workstation.

The most surprising detail is the research artifact shipped alongside. It shares the architecture but was trained to accept a variable number of passes through that loop. Setting one or two passes instead of three makes inference much faster, the authors say, at some cost in quality — a speed dial for local users. The team says it tested the variant and that it works, but has not evaluated it in depth.

On benchmarks, the source claims state-of-the-art status in its class on the Artificial Analysis Intelligence Index v5 and says the model leads four named rivals — Gemma 4 31B, Muse Glimmer, Qwen3.8-27B and Nemotron 3.5 Super VL — on 32 of 48 tests spanning coding, agentic tool use, vision, reasoning and safety. It leads on FrontierCode v2.0, ties Qwen3.8-27B on Terminal-Bench 4.0 and trails it on AutomationBench-AA. The authors describe their own comparisons as "vibe tests" and call for more validation, so treat the rankings as a

promising signal rather than a settled result. The demonstrations — a Van Gogh painting recreated as an animated shader, a hand-drawn wireframe turned into a working expense app — show what a small local model can now attempt, not what it reliably delivers.

KEY FACTS

TOTAL PARAMETERS **~26B**	ACTIVE PARAMETERS PER TOKEN **3B**
CONTEXT WINDOW **1M tokens**	BENCHMARKS LED **32 of 48**
RELEASED BY **Thinking Machines Lab, on Hugging Face**	LOCAL FORMATS **4-bit (IQ4_NL) and 8-bit (Q8_0)**

> **WHY IT MATTERS**
>
> Teams that want a capable multimodal model without sending data to a cloud API now have another option to test, and one small enough for a laptop or workstation. The benchmark claims are self-reported and lightly validated, so buyers should run their own checks before switching.

Read the original at huggingface.co →

---

03 **MEDIUM IMPACT** MIT Technology Review

METR finds AI coding gains do not guarantee good judgment

> AI models keep getting better at coding and maths at a steady exponential clip, but they still fall short on work that requires judgment, according to a new report from the evaluation nonprofit METR.

METR, a nonprofit that tests AI systems, published an update to its widely shared "time horizon" chart, which measures how long a task a model can complete, judged by how long the same task takes a human. The original plot found that the length of tasks frontier models could finish doubled every four months. The new version adds harder and more open-ended tasks, including ones meant to resemble the work of AI research itself.

The gap is the story. Models have become excellent at "verifiable" tasks — coding and maths, where a machine can automatically check whether an answer is right. Researchers train them on those tasks through reinforcement learning: attempt the task many times, strengthen what worked, weaken what failed. That loop is hard to run on work where success is fuzzy or shows up late, such as picking a research direction. In one new task, models were asked to design AI evaluations METR could use; human employees scored the results, and not one was good enough to be considered useful. On scoring philosophical arguments, models also trailed humans.

The most surprising detail: even the best models now beat almost every human at software engineering, yet they cannot produce a single usable research evaluation. Some of the newest models, such as Opus 5.5, reach what METR calls "borderline above the bar" on evaluation design — not usable, but promising. A year ago every model was "clearly below the bar."

The caveat is measurement. In the original plot, humans actually did each task to time it. In the new report, completion times are mostly estimates, because paying people for tasks that might take hundreds of hours is too

expensive. For the philosophy task, one person rated most of the arguments. METR's Matt Putz said trends on judgment-heavy tasks will likely never be as clean as for easy-to-specify ones. The update is a step toward testing what matters for automating research, not an answer to whether that is coming.

KEY FACTS

TASK LENGTH DOUBLING
Every four months

AI-DESIGNED EVALUATIONS JUDGED USEFUL BY METR STAFF
None

BEST CURRENT RATING ON EVALUATION DESIGN
"Borderline above the bar"

RATING A YEAR AGO
"Clearly below the bar"

CLAIMED AI SPEED-UP OF AI PROGRESS AT ANTHROPIC
Factor of two

WHY IT MATTERS

Companies selling AI on the promise of automating research and development should expect buyers to ask for evidence on open-ended work, not just coding benchmarks. For everyone else, the practical read is that today's models are strongest where answers can be checked automatically.

Read the original at MIT Technology Review →

---

04 HIGH IMPACT arXiv.org

AI bug-hunting review finds more flaws, but no faster fixes

> AI systems are now finding and fixing real software flaws on their own, but the organisations that own that software are not patching them any faster, according to a review of published evidence.

The paper, a systematisation of knowledge — an academic review that pulls scattered results into one comparable picture — rates AI agents' vulnerability work on a scale borrowed from self-driving cars, from no automation up to full autonomy. It then checks each published claim for whether it is correct, whether the tools are available, and how strong the evidence behind it is. The authors conclude that agents discovered, validated and patched vulnerabilities in real-world software without human intervention before the restricted frontier models of 2026, and that open-weight models — ones anyone can download and run — later reached comparable autonomy for finding and validating bugs.

The consequences show up in a decade of public vulnerability records. In 2026 the number of disclosed vulnerabilities more than doubled, with a growing share rated critical or high. Severe bugs that had survived ten years or more in widely used code were reported several times more often than before. That is the detail worth repeating to a colleague: the long-lived, deeply buried flaws that security teams assumed were already found are now surfacing in volume.

What has not changed is speed. The paper measures the gaps between the stages of a vulnerability's life — from report to fix, and from disclosure to attack — and finds no noticeable shift so far. Finding flaws faster has not made anyone fix them faster, and it has not visibly shortened the window between a flaw becoming public and being exploited. For businesses running software, that is the practical finding: the queue of known problems is

growing while the time to close them stays flat. For investors, it points at where the bottleneck now sits — in patching and maintenance work, not in detection.

The genuinely new contribution is the framework itself. Vendor reports and research papers on AI security capability have rarely been comparable, and the authors grade the proposed mitigations against an explicit rubric before drawing guidance for each group of stakeholders. The finding that autonomy arrived before the restricted 2026 frontier models, and that open-weight models then caught up on discovery and validation, is the part that changes assumptions about who can run this kind of tooling.

KEY FACTS

CHANGE IN DISCLOSED VULNERABILITIES, 2026
More than doubled

SEVERE BUGS SURVIVING A DECADE OR MORE
Reported several times more often than before

TIME FROM REPORT TO FIX
No noticeable shift so far

TIME FROM DISCLOSURE TO ATTACK
No noticeable shift so far

OPEN-WEIGHT MODELS
Reached comparable autonomy for discovery and validation

SUBMITTED
19 Nov 2026

> **WHY IT MATTERS**
>
> If AI is finding more flaws but fixes are not arriving faster, the pressure falls on patching and maintenance capacity — the part of security that costs staff time. Businesses should expect a longer list of known vulnerabilities, not a shorter one.

Read the original at arXiv.org →

SECTION 2 OF 3

AI Tools & Ecosystem

4 stories · 2 high · 2 medium

05 MEDIUM IMPACT huggingface.co

Hugging Face cuts coding-agent token use with custom training

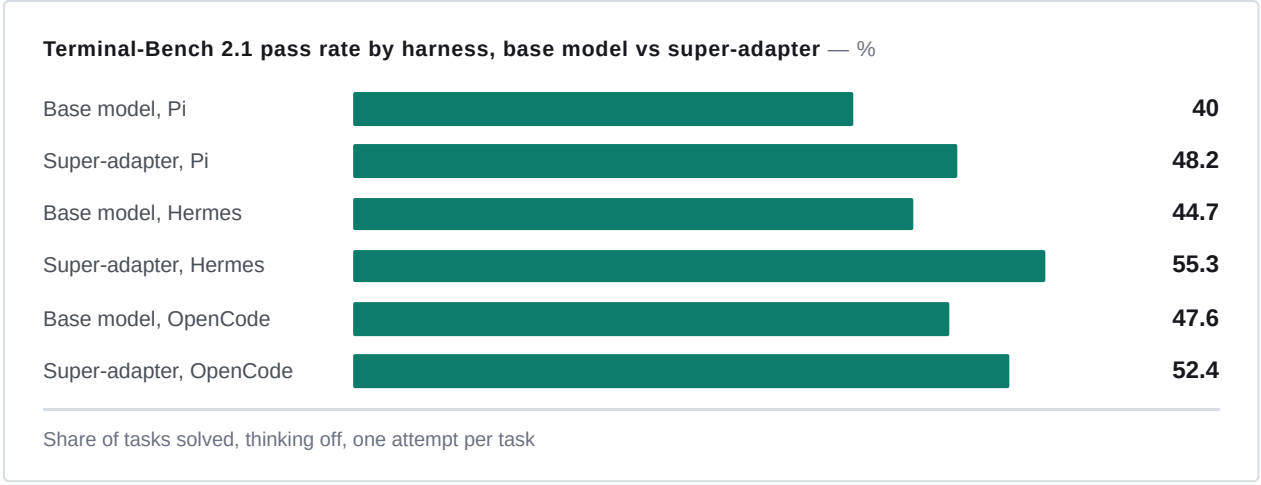
> Hugging Face has published a reusable recipe for training a coding model to work better inside the specific agent software a team already uses, and released the adapters, datasets and training code alongside it.

The company took Qwen3.8-27B, an open model, and trained a separate add-on adapter — a LoRA, a small set of extra weights rather than a retrained model — for each of three agent harnesses: Pi, OpenCode and Hermes Agent. A harness is the software wrapper that lets a model call tools, read files and run commands; the same model behaves differently in each. Under Pi, the model's score on Terminal-Bench 2.1, a test of real software-engineering tasks, rose from 39% to 55%, while average token use per task fell from 2.9K to 1.5K. The gain comes mostly from the model reasoning less, calling fewer tools and getting it right first time.

The practical pitch is cost. On tasks both the base model and the adapted version solved, the specialists made 53–72% fewer tool calls on the hard SWE-bench Pro set and 53–70% fewer on Terminal-Bench. Every tool call is a paid round trip to a model, so fewer calls is a smaller bill and a faster answer. Anyone running a coding agent at volume — a software team, a platform vendor, a startup metering usage — is the audience. The base model and the adapters are open, so the recipe can be copied rather than licensed.

The most surprising result is the merged "super-adapter". The three adapters turned out to point in almost unrelated directions — the source puts the similarity between any two at about 0.01–0.02 — so combining them cost little. A single adapter at half strength beat the base model under all three harnesses, by 6 to 18 points on SWE-bench Pro and 5 to 10 on Terminal-Bench, with a 53–73% cut in tool calls. It also improved on the base model's "thinking" mode in five of six harness–benchmark pairs, despite never being trained with thinking on.

What is new is the packaging, not the idea. Training a model inside its own harness has been shown before, by Hugging Face and others; here it is done at 27B on harder tasks, across three harnesses, with the task-generation, validation and filtering pipeline published. Of 1,127 candidate tasks, 858 passed validation and only 171 were kept as useful for training — the rest were either broken or too easy or too hard to teach anything. The limits are stated plainly: on unrelated business-automation tasks the adapters stayed close to the base model, and the authors call what the model actually learned an open question.



KEY FACTS

SCORE UNDER PI HARNESS, BEFORE
39% on Terminal-Bench 2.1

SCORE UNDER PI HARNESS, AFTER ADAPTER
55%

TOKENS PER TASK, BEFORE AND AFTER
2.9K to 1.5K

FEWER TOOL CALLS ON SOLVED TASKS
53–73%

TASKS KEPT FOR TRAINING
171 of 442 scored; 871 in the released dataset

TRAINING COMPUTE
2x B200 GPUs per adapter; 167M tokens trained

WHY IT MATTERS

Teams paying per tool call for coding agents can cut token and call volume by half or more by training a cheap adapter on their own tasks and harness, instead of switching to a bigger model. The catch is that it takes engineering work and a supply of validated tasks, which Hugging Face has now published.

Read the original at huggingface.co →

---

06 HIGH IMPACT Databricks

Databricks index shows identical coding scores can hide huge costs

> Databricks published an index of how 10,000-plus companies actually buy and use AI models, and its coding benchmarks show two models with identical scores where one costs 32 times more per task.

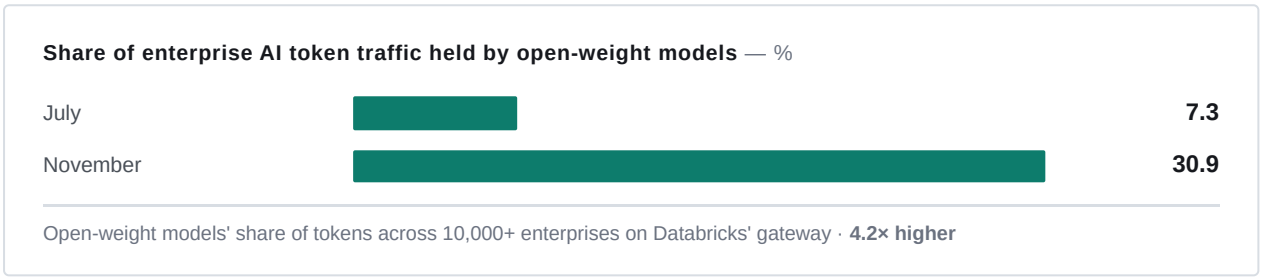
The data company Databricks, which sells the software layer many enterprises use to access AI models, has made public a weekly-updated index built from anonymised usage across more than 10,000 customer companies, plus coding tests drawn from its own engineers' real work. The dataset covers July 20 to November 7, 2026. It is the first time Databricks has opened its internal evaluation suite, and it answers a question customers keep asking: which model is worth paying for.

The headline finding is that benchmark scores alone mislead. In Databricks' coding test, GPT-6 Luna and Claude Opus 5.5 both completed 89.1% of tasks, but Luna averaged $0.05 a task against $1.62 for Opus 5.5. Cost is not just the sticker price of tokens: Gemini 3.8 Flash has a low token price yet averaged 140 turns per task and ended up among the most expensive at $11.97. A model that needs more attempts or longer answers burns money even when each word is cheap. At the top end, Claude Fable 5.1 scored 96.0% at $7.47 a task

against Muse Wonder's 94.9% at $1.53 — 1.1 points for nearly five times the price. Fable 5.1's share of proprietary tokens fell from 5.5% to 0.6% in six weeks.

The index also records a rapid loosening of vendor loyalty. In July about 64% of enterprises on Databricks' gateway used a single model provider; by November 2 that was 28%. Open-weight models — ones a company can host itself rather than rent from a lab — rose from 7.3% to 30.9% of enterprise token traffic in under four months, and three now sit in the coding top ten, including DeepSeek V4.1 and GLM 5.3 at a 93.1% pass rate. Quality still moves share fast: after Claude Opus 5.5 launched, it went from 7.2% to 37.9% of proprietary tokens in six weeks, lifting Anthropic from 57.6% to 68.8%.

Spending is lopsided. The median user spent $2.49 a week on AI; the 99th-percentile user spent $699.74, about 281 times more. Median spend more than doubled from $1.10 in six weeks while the top deciles stayed flat, so most of the bill still comes from a few heavy users. Databricks says it stayed on an older Opus model internally after replaying its own workload traces showed the upgrade was not worth the cost.



KEY FACTS

DATA COVERS
July 20 to November 7, 2026

COMPANIES IN USAGE DATA
10,000+

COST GAP AT EQUAL SCORE
$0.05 vs $1.62 per task (GPT-6 Luna vs Claude Opus 5.5, both 89.1%)

OPEN-WEIGHT SHARE OF ENTERPRISE TOKENS
7.3% in July to 30.9% in November

SINGLE-PROVIDER ENTERPRISES
64% in July, 28% by November 2

MEDIAN VS 99TH-PERCENTILE WEEKLY AI SPEND
$2.49 vs $699.74

> **WHY IT MATTERS**
>
> Buying AI on benchmark scores can cost a company tens of times more than it needs to for the same results; the useful measure is cost per finished task on your own work, not token price or a leaderboard rank.

Read the original at Databricks →

---

07 MEDIUM IMPACT Source

Microsoft and 40 organizations back verifiable AI-agent actions

> Microsoft has launched an AI Transparency Initiative with more than 40 organizations to build a shared, open way of recording and checking what AI models and agents do inside a company's systems.

Announced by chairman and chief executive Satya Nadella at Microsoft's Ignite conference on Tuesday, the effort pairs Microsoft with Accenture, Adobe, KPMG, Mastercard and Volvo Cars, among others, to develop

standards for AI systems that are accountable, independently verifiable and safe by design, whatever the model or provider. Microsoft will also fold the work into Microsoft AI, Microsoft Foundry and Microsoft Copilot, so customers can feed the resulting evidence into their own security and compliance tools.

The problem it addresses is an evidence gap. Companies now wire AI agents into data, tools, credentials and infrastructure, and those agents can act without a human approving every step. Existing controls can watch much of the activity around a model but cannot see inside it, and cannot independently check its account of its own behaviour. Microsoft's security executive vice president Hayete Gallot said organisations need greater transparency into the model itself for security controls to work. Adobe's chief security officer, Aanchal Gupta, said security teams need clear visibility into what agents access and what actions they take.

Microsoft's framing is blunt: an agent with access to data, tools and credentials carries the risk profile of a privileged insider but operates at machine speed and scale. Accountability stays with the organisation, the company says, and blaming the AI is not an answer. The proposal rests on three verbs: see what the AI was asked and authorised to do, which model and runtime were used and what it touched; verify that activity against independent records rather than the model's own report; and control what it is permitted to do before consequential actions occur. The evidence should stay with the customer when models or providers change.

The initiative has three parts: expanded participation in the National Institute of Standards and Technology's SI Agent Standards Initiative, collaboration with the 40-plus organisations on open standards, and product integration. What is new is the push for cross-provider verification, including closed and open-weight models, so developer teams can swap models without stripping IT and compliance teams of the ability to govern and audit them. What is not yet here is a standard: the announcement sets out principles and a work programme, not a finished specification or a date for one.

KEY FACTS

ORGANIZATIONS INVOLVED
More than 40

ANNOUNCED
Nov. 17, 2026, at Microsoft Ignite

NAMED PARTNERS
Accenture, Adobe, KPMG, Mastercard, Volvo Cars

MICROSOFT PRODUCTS AFFECTED
Microsoft AI, Microsoft Foundry, Microsoft Copilot

STANDARDS BODY
NIST SI Agent Standards Initiative

> **WHY IT MATTERS**
>
> Companies deploying agents that touch money, customer records or infrastructure should expect auditors and regulators to ask for independent evidence of what those agents did — and this group is trying to make that evidence portable across vendors rather than tied to one model maker.

Read the original at Source →

---

08 **HIGH IMPACT** Salesforce

Copilotforce brings Salesforce customer records into Microsoft Copilot

> Salesforce and Microsoft announced Copilotforce at Microsoft Ignite, a partnership that puts Salesforce customer records and sales workflows inside Microsoft Copilot, Teams and Fabric.

The deal, announced November 17, 2026, extends Salesforce's Sales Cloud into Microsoft's productivity software, where sellers already spend their day. In Copilot, sales staff can ask plain-language questions about an account and get answers built from Salesforce data combined with Teams meeting transcripts, Outlook emails and shared documents. They can also take supported Salesforce actions without leaving Copilot. A separate Teams app, already in pilot since Dreamforce, shows pipeline and quota progress and lets sellers update opportunity details that sync back to Salesforce.

The second piece is data plumbing. Salesforce Data 360 and Microsoft Fabric will connect through what the companies call bidirectional zero-copy access: organizations can query Fabric data inside Data 360 and share Data 360 customer intelligence into Microsoft's OneLake storage layer without building duplicate datasets or maintaining the custom pipelines that normally move data between systems. Data federation is generally available; the sharing direction is in public preview.

The most concrete proof point is Accenture, which is both design and implementation partner and customer zero — the first company to run the integrations in its own business. Accenture has rolled out Microsoft Copilot to more than 740,000 people and runs Salesforce with 55,000 active users, giving the partnership an unusually large internal test bed. Salesforce's Miguel Milano framed the bet as reaching "the largest surface in enterprise software today"; Microsoft commercial chief Judson Althoff stressed that customer data stays under existing permissions and audit trails, with any write-back to Salesforce requiring explicit user confirmation.

What is genuinely new is the combination: Salesforce context inside Copilot, a Teams app moving from pilot toward wider use, and a live data link between the two vendors' clouds. What is not yet available broadly is the headline Copilot integration, which is in public preview, or the Data 360 sharing feature, also in beta. Salesforce's own release notes tell customers to base purchasing decisions on capabilities that are commercially available today. For buyers, the practical question is whether the previews hold up inside their own permission structures before they commit budget.

KEY FACTS

ANNOUNCED
November 17, 2026, at Microsoft Ignite

SALESFORCE IN MICROSOFT COPILOT
Public preview (beta)

SALESFORCE IN TEAMS
Pilot

DATA FEDERATION
Generally available

DATA 360 DATA SHARING
Public preview (beta)

ACCENTURE COPILOT ROLLOUT
More than 740,000 people

WHY IT MATTERS

Companies that already pay for both Salesforce and Microsoft 365 may get more use out of each without a migration project, but the main Copilot integration is still a preview, so the value is not yet guaranteed. The Accenture deployment shows the scale of the bet — and the size of the test.

Read the original at Salesforce →

SECTION 3 OF 3

AI Applications & Industry

4 stories 1 high 3 medium

09 HIGH IMPACT Ars Technica

Two million bait secrets expose leaks across AI services

> A security researcher planted two million uniquely trackable secrets in prompts sent through the network of companies that route and resell AI model access, and found that sensitive data leaked at nearly every link in the chain.

HD Moore, chief research officer at the security firm Dragos, built a set of working Amazon Web Services access keys tied to empty accounts, SSH passwords, private web addresses and git repository references. None had been used before, and he controlled everything behind them. Each secret went into exactly one prompt to one provider, so any reappearance identified the party responsible. He called them canaries, after the birds sent into coal mines.

The results, published Thursday, describe an ecosystem that leaks like a sieve. At least three AI providers were compromised by attackers who used their access to take prompts they received. In one, every request sent during a few short windows in late October was stolen, including requests the provider rejected before a model ever ran — meaning the attacker was copying traffic at the provider's front door. In the other two, some stolen requests carried Zero Data Retention flags, a setting meant to stop data being kept. The stolen keys were used tens of thousands of times in AWS calls, with intruders hunting through storage buckets and secrets managers for more credentials. Secrets sent to two of the hacked providers were used together almost immediately from the same rented server, suggesting one adversary.

Separately, several resellers in the inference marketplace logged prompts and passed them on. They drew traffic with low prices, switched on request logging and waited. Moore caught them by sending the same request through different sellers and watching which one leaked. Some providers also ignored privacy settings outright: one Chinese AI lab's systems fetched secret web addresses from more than half the prompts sent to its models between July and October, including about two-thirds of those marked no data collection, typically two or three days later. Several US hosting providers resell that lab's models, and prompts sent to them triggered the same scraper — so a customer who chose a US provider to keep data in the country did not get what they paid for.

The mechanics explain the exposure. Prompts are encrypted in transit but decrypted at every hop inside the reseller network, because each router must read the text to pick a model. Each player is likely to log and store it. Moore's advice: host models on your own servers if you can afford it; otherwise buy directly from a first-tier provider with one contract and a privacy policy, and choose the upstream provider yourself rather than accepting a default pool.

KEY FACTS

PROMPTS SEEDED WITH TRACKABLE SECRETS
Nearly 2 million

PROVIDERS BREACHED
At least three

STOLEN KEYS USED IN AWS CALLS
Tens of thousands of times

KEYS REUSED BY ONE ATTACKER
A couple hundred

SECRET URLS FETCHED BY ONE LAB
More than half of prompts, July to October

DELAY BEFORE SCRAPER APPEARED
Two or three days

> WHY IT MATTERS
>
> Anyone pasting client data, code, credentials or internal documents into an AI tool may be handing them to companies they never chose, some of which keep or resell them. If your business relies on AI apps, ask which provider actually serves the request and whether self-hosting is affordable.

Read the original at Ars Technica →

---

10 MEDIUM IMPACT TechCrunch

Lenders resist financing SpaceX chips that cannot be repossessed

> Lenders behind SpaceX's $40 billion chip financing asked that none of the chips be used for the company's orbital data center plans, according to the Financial Times, because hardware in orbit cannot be seized if a borrower defaults.

The data center boom runs on debt. Cloud providers borrow to build, using the chips themselves as collateral: if a borrower defaults, lenders take the hardware and resell it. That backstop is what makes chip financing cheap. Older chips hold value — research firm Compute Credit Index says some H100s sell for about half their 2023 purchase price, and some H200 configurations fetch what they did when introduced in 2024.

A chip installed in a satellite 300 miles up, travelling at 17,500 miles per hour, cannot be repossessed. That is the problem lenders raised over SpaceX's $40 billion chip financing, asking that none of the purchased chips feed the company's orbital data center plans. It is not clear whether SpaceX agreed. The company has room to wait: its ground-based compute build-out is already lucrative, and on its latest earnings call chief executive Elon Musk said he was moving engineers off the rocket team to speed it up, including work on casting turbine blades for generators faster than existing suppliers.

The constraint is not only money. Finding land and power for terrestrial data centers is getting harder, with projects facing delays and local opposition, while the market waits to see whether compute demand holds and whether buyers can turn it into revenue. Musk has said since the start of 2026 that space will be the cheapest place to run data centers within two to three years — a claim that depends on SpaceX flying Starship in a fully reusable configuration. In its most recent test the booster stage was not recovered, though the orbital stage landed back at the Texas launch site. When the second stage flies again, and what refurbishing it costs, is unknown.

SpaceX and rivals including StarCloud, Cowboy Space Corporation, Sophia Space and Orbital all still have to buy chips, and are likely to pay more than ground-based operators. Google, deploying its own TPU chips in orbit through Project Suncatcher, may benefit. SpaceX and Tesla are jointly building a U.S. chip foundry, Terafab, but it has not broken ground and would take years and billions of dollars to reach volume production.

KEY FACTS

SPACEX CHIP FINANCING DEAL
$40bn

ORBIT ALTITUDE CITED
300 miles

ORBITAL SPEED CITED
17,500 mph

USED H100 RESALE PRICE
About half of 2023 cost basis

MUSK'S CLAIM
Space cheapest for data centers in 2-3 years

JOINT SPACEX-TESLA FOUNDRY
Terafab, not yet broken ground

> WHY IT MATTERS
>
> The collateral that makes data center debt cheap does not exist in orbit, so space compute will likely cost more to finance than the ground variety — a real brake on the orbital data center pitch, and a quiet edge for whoever already owns chips on Earth.

Read the original at TechCrunch →

---

11 MEDIUM IMPACT TechCrunch

Spotify's Daniel Ek backs a simulation-first fusion startup

> Spotify founder Daniel Ek has put money behind Europa Fusion, a secretive UK fusion power startup founded in February by an artificial intelligence researcher who argues the industry spends too much on hardware and too little on simulation.

The company was set up by Nantas Nardelli, who has worked at Google DeepMind, Facebook AI Research and Microsoft Research Cambridge, and was most recently head of machine learning at Pacific Fusion, a California fusion startup. Bloomberg first reported Ek's backing; TechCrunch reported the founder's identity and approach, citing people it did not identify because the information is not public.

The split with Pacific Fusion is the substance of the story. Pacific Fusion recently broke ground on a demonstration facility in New Mexico and, like most fusion startups of its size, leans heavily on physical devices to test its theories and sharpen its simulations. A source familiar with Nardelli's thinking told TechCrunch he did not agree with that hardware-first focus, and thinks fusion companies should put more into simulation and data instead. Europa Fusion will still pursue the same pulsed-power route: reactors that fire huge, precisely timed surges of electrical current to compress fuel targets until the atoms inside fuse and release energy.

That makes this a bet about where the money goes before anything is built. Fusion has long promised carbon-free power from nearly limitless fuel, and no one has yet built a plant that supplies electricity to a grid. If it works commercially, it could also ease concerns about the electricity that AI data centres consume — a neat loop, given that an AI researcher is the one making the pitch. Nardelli's own explanation for leaving the field is on his website: "My original background is in electronics and control engineering," he wrote. "I used to build robots for a living. One day I'll be back to that, but it's time to fix climate change for good."

What is genuinely new is the argument, not a machine. No funding figure, reactor design or timeline has been disclosed, so the claim that simulation can substitute for some physical testing remains untested in public. For investors, the signal is that a well-known consumer-tech founder is willing to fund a pre-hardware fusion

concept; for the fusion sector, it is a challenge to the assumption that building devices is the only way to make progress.

KEY FACTS

INVESTOR
Daniel Ek, Spotify founder

COMPANY
Europa Fusion, UK-based

FOUNDED
February

FOUNDER
Nantas Nardelli

FOUNDER'S PREVIOUS EMPLOYERS
Google DeepMind, Facebook AI Research, Microsoft Research Cambridge, Pacific Fusion

APPROACH
Pulsed-power fusion, simulation- and data-heavy

> **WHY IT MATTERS**
>
> Fusion investors and founders now have a visible test case for whether software and simulation can cut the cost and time of developing a power plant, rather than pouring money into test machines first. For anyone watching AI's energy footprint, a well-funded fusion effort is one more bet on supplying the electricity that data centres will need.

Read the original at TechCrunch →

---

12 MEDIUM IMPACT TechCrunch

Google commits over $500 million to super-hot geothermal power

> Google has agreed to pay more than $500 million for 50 megawatts of electricity from Quaise Energy, a startup that plans to drill three miles down into super-hot rock in central Oregon.

The deal, announced Tuesday, makes Google a customer of a technology that has never been delivered commercially. Quaise intends to reach the heat by starting with an ordinary drill bit, then switching at about 2.5 miles to a gyrotron — a device that fires millimeter-wave radiation at the rock and turns it to ash. Metal bits wear out quickly at the temperatures involved, and the company's bet is that vaporising the rock is cheaper than replacing bits.

Google is paying a premium per megawatt for power Quaise has not yet produced. The company says that fits its pattern of signing up early with potential suppliers; it was both an investor in and an early customer of Fervo Energy, which went public earlier this year and gave public markets their first taste of enhanced geothermal. The prize is size: the U.S. has up to 57 terawatts of geothermal generating capacity, according to the source, far more than the next wave of AI data centers needs. A single super-hot rock well could produce 15 megawatts or more, several times a conventional well.

The most surprising detail is the physics being exploited. At extreme depth, heat and pressure turn ordinary water into a supercritical fluid — part gas, part liquid — that carries far more energy than steam. Drilling is the cost that decides whether this works: it accounts for roughly a quarter to a third of a geothermal plant's total cost, so getting more power from each well is the whole game.

Quaise has raised more than $270 million from investors including Prelude Ventures, Collaborative Fund and Engine Ventures, with Vinod Khosla as an angel. Rival Mazama Energy, backed by Khosla Ventures, Doerr

Capital and Bill Gates, has raised $171 million and said Monday that its Carbon Chill system — using carbon dioxide instead of water-based mud to cool the bit — works. Chief executive Sriram Vasantharajan told TechCrunch its second well was drilled 80% faster than its first, and the cooling tech cut drilling time a further 20%. XGS Energy, meanwhile, received a Department of Energy grant last week. None of this is proven at commercial scale, and the risk that the techniques don't pay off remains.

KEY FACTS

GOOGLE COMMITMENT
More than $500 million

POWER SECURED
50 megawatts

WELL DEPTH
Three miles

QUAISE FUNDING RAISED
More than $270 million

MAZAMA FUNDING RAISED
$171 million

U.S. GEOTHERMAL POTENTIAL
Up to 57 terawatts

> **WHY IT MATTERS**
>
> Tech companies are contracting for power years before it exists, which locks in costs and suppliers for data-center buildouts. For investors, the geothermal sector now has a public comparable in Fervo and a large anchor customer in Google, but no commercial super-hot well yet.

Read the original at TechCrunch →

Sources

01 Claude helps claim a Riemann proof as a human rival publishes
https://techcrunch.com/2026/11/20/ai-claimed-a-solution-to-the-167-year-old-riemann-hypothesis-and-a-human-did-too/

02 Inkling 2 Mini handles images and text on local hardware
https://huggingface.co/blog/inkling2mini

03 METR finds AI coding gains do not guarantee good judgment
https://www.technologyreview.com/2026/11/20/1149399/the-most-misunderstood-graph-in-ai-just-got-a-major-update/

04 AI bug-hunting review finds more flaws, but no faster fixes
https://arxiv.org/abs/2611.26601

05 Hugging Face cuts coding-agent token use with custom training
https://huggingface.co/blog/harness-adapters-at-scale

06 Databricks index shows identical coding scores can hide huge costs
https://www.databricks.com/blog/introducing-databricks-ai-index-state-enterprise-ai-tokenomics

07 Microsoft and 40 organizations back verifiable AI-agent actions
https://news.microsoft.com/source/2026/11/17/microsoft-and-organizations-across-industries-launch-ai-transparency-initiative-to-strengthen-enterprise-trust-in-ai-models-and-agents

08 Copilotforce brings Salesforce customer records into Microsoft Copilot
https://www.salesforce.com/news/stories/copilotforce-announcement

09 Two million bait secrets expose leaks across AI services
https://arstechnica.com/security/2026/11/a-leaky-ai-supply-chain-is-exposing-secret-user-data-at-an-alarming-rate/

10 Lenders resist financing SpaceX chips that cannot be repossessed
https://techcrunch.com/2026/11/20/chips-can-survive-the-space-environment-can-their-financing/

11 Spotify’s Daniel Ek backs a simulation-first fusion startup
https://techcrunch.com/2026/11/20/daniel-ek-backs-ai-researchers-fusion-power-startup-heres-what-theyre-building/

12 Google commits over $500 million to super-hot geothermal power
https://techcrunch.com/2026/11/20/super-hot-rock-geothermal-could-end-data-center-energy-woes

---

About this document. Every story in the 21 November 2026 New Horizon AI Digest, reported at length. Each entry is written from the publisher's own article text; where a source could not be retrieved the entry is explicitly marked and kept short rather than padded.

Images and licensing. Figures are used only where the source licence permits redistribution, and are credited in the caption. Publisher artwork is not reproduced. All titles link to the original publication.

Vocabulary. Common AI terms are defined at new-horizon.tech/glossary/.

Generated 2026-11-21T06:37:27Z · model deepseek-v4.1-flash:cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.