NEW HORIZON · AI DIGEST

The Long Read

19 November 2026

12	3	4403	4
STORIES	SECTIONS	WORDS	HIGH IMPACT



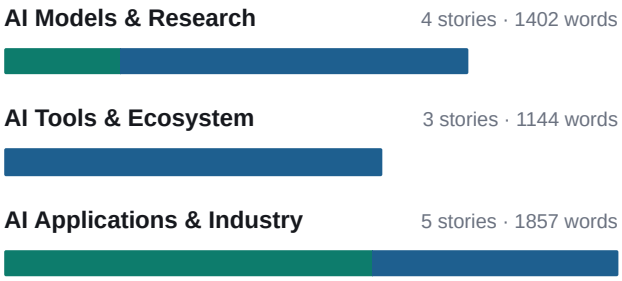
4 high impact 8 medium impact spoke length = depth of coverage

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 · 4403 words · 3 sections · 2 charted





AI Models & Research

01 Google’s Gemini 4 Neon cuts task costs by 32%

02 Claude solves two open problems in black-hole physics

03 Rumble releases a compact speech model for 40+ languages

04 Google opens Genie 4’s interactive worlds to Ultra subscribers

AI Tools & Ecosystem

05 Hugging Face runs a giant open model on a 32GB Mac

06 Snapchat lets coding assistants build and publish camera effects

07 Ollama shows why small prompt changes inflate agent costs

AI Applications & Industry

08 Nadella says Microsoft’s 60,000 buyout offers reflect an AI work redesign

09 Citi makes customers liable for AI-agent losses, even after hacking

10 Anthropic reportedly prices IPO at $83 for a $1.5 trillion valuation

11 Humyn Labs raises $47 million to record robot-training data

12 Osmo says its AI nose matches people at distinguishing smells

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 HIGH IMPACT Google

Google’s Gemini 4 Neon cuts task costs by 32%

> Google released Gemini 4 Neon, a model it says matches frontier-level reasoning while costing 32% less per task than its own larger Carbon model.

Neon is the middle child of Google's Gemini 4 family: Carbon handles the hardest, longest projects, Hydrogen handles quick everyday jobs, and Neon is pitched at the high-volume work in between — coding, analysis, spreadsheets, research. Google says it scores 61 on the Artificial Analysis Intelligence Index at its highest thinking setting, third overall behind larger, slower models, and ranks second on the Vals Index.

The pricing is the story for anyone paying an AI bill. Neon costs $3 per million input tokens and $15 per million output tokens, with cached input at $0.15 per million — a 95% discount on reused material. Google says that at its default setting it sits on the cost-efficiency frontier on two independent indexes, roughly 50% to 80% cheaper per task than top competing frontier models, and at its highest setting still costs at least 36% less per task than rivals running at their maximum. For a company running agents across thousands of documents a day, that gap compounds fast.

The most surprising claim is about restraint. Google reports Neon refused just 1.98% of tasks on one coding benchmark and 0% on a science benchmark, while matching Carbon's 0.0% cheating rate and taking shortcuts on 2.45% of tasks in its own coding evaluation — the lowest of any Gemini 4 model. It also says Neon resisted every attack in an external test where malicious instructions were planted in web pages, emails and code it reads. That combination — cheap, capable, and hard to hijack — is what enterprise buyers actually ask about.

A separate cyber version, built to write and harden secure code and to find and patch vulnerabilities, goes to vetted defenders through Google's Fairwind Program in the coming days. Neon itself is available now through the Gemini API, Google AI Studio, Gemini for business, and to paid Pro and Ultra subscribers in the Gemini app and Antigravity. The incremental part is the price; the new part is a mid-tier model that Google claims is safe enough to leave running unattended.

KEY FACTS

INPUT PRICE
$3 per million tokens

OUTPUT PRICE
$15 per million tokens

CACHED INPUT DISCOUNT
95% ($0.15 per million tokens)

COST PER TASK VS CARBON
32% lower at default setting

CONTEXT WINDOW
1 million tokens

CYBER VERSION ACCESS
Fairwind Program, rolling out in coming days

WHY IT MATTERS

Cheaper per-task reasoning changes the maths on automating high-volume work: tasks that were too expensive to hand to a model last quarter may now pay for themselves. Buyers should re-run their cost estimates rather than assume last year's prices.

Read the original at Google →

02 MEDIUM IMPACT www.anthropic.com

Claude solves two open problems in black-hole physics

> Anthropic says its Claude model has solved two long-standing problems in theoretical physics, one of them a calculation physicists set out in 2003 and never finished.

The results, posted by Anthropic on its research site, concern black holes and the search for a theory of quantum gravity — the framework that would reconcile general relativity, which describes stars and gravity, with quantum mechanics, which describes atoms. The two theories are mathematically incompatible, and black holes are one of the few places where both matter.

The first problem dates to 1983, when Stephen Hawking and Don Page showed that a model universe heated past a certain temperature switches abruptly from a gas of particles to a black hole, like water boiling. Their calculation relied on general relativity, which breaks down at small scales. In 1999 physicists found a similar transition on the quantum side but could not tell whether it was abrupt or gradual; in 2003 a group set out the calculation that would settle it, but it was too complex to finish. Claude ran a group of copies of itself, working out almost every piece independently and checking against known results, and found the transition is abrupt — suggesting black hole formation is built into the underlying quantum theory, not just relativity.

The second problem concerns a simplified two-dimensional model universe. Since 2009 physicists have used a method called bootstrapping to set a ceiling on how heavy the lightest black hole could be, ruling out any theory that breaks it. Numerical results hinted at a lower ceiling than anyone could prove. Claude noticed a pattern in those numbers that physicists had missed, borrowed a method from elsewhere in maths, and calculated the precise ceiling for the simplest version of the technique, proving it could not be lowered.

The detail a reader would repeat: the work began as a side project. In August an Anthropic researcher, Hongbin Chen, asked Claude to list interesting open problems in theoretical physics; it returned 22 candidates, and separate agents ran in the background for about six weeks with no special setup. One agent had reached the Hawking-Page answer before a physicist who had tackled the same problem independently. The results were formally verified in Lean, a computerized proof checker, and a physicist checked every main result. Neither finding describes our own universe or has practical applications.

KEY FACTS

PROBLEMS SOLVED
Two open problems in theoretical physics

CALCULATION SET UP
2003, never completed until now

CANDIDATE PROBLEMS CLAUDE PROPOSED
22

TIME AGENTS RAN IN BACKGROUND
About six weeks

FAILED PARALLEL ATTEMPTS BEFORE SUCCESS
61

VERIFICATION
Formally verified in Lean, a computerized proof checker

> WHY IT MATTERS
>
> The results are theoretical and have no commercial use, but they are a concrete test of whether AI agents can grind through research problems humans abandoned as too laborious — a question that matters to anyone betting on AI as a tool for science and engineering.

Read the original at www.anthropic.com →

---

03 MEDIUM IMPACT huggingface.co

Rumble releases a compact speech model for 40+ languages

> Rumble has released Rumble-2-Flash, an open-source speech recognition model that handles more than 40 languages and, on a benchmark of real call-centre audio, performs about as well as a model 60 times larger.

The model is published under the Apache 2.0 licence, which lets companies use it commercially without paying for a licence or asking permission. It is small enough to run on ordinary hardware, and it works out of the box with two widely used software tools for running AI models, so a team does not need custom code to deploy it. Rumble also offers it through its own API.

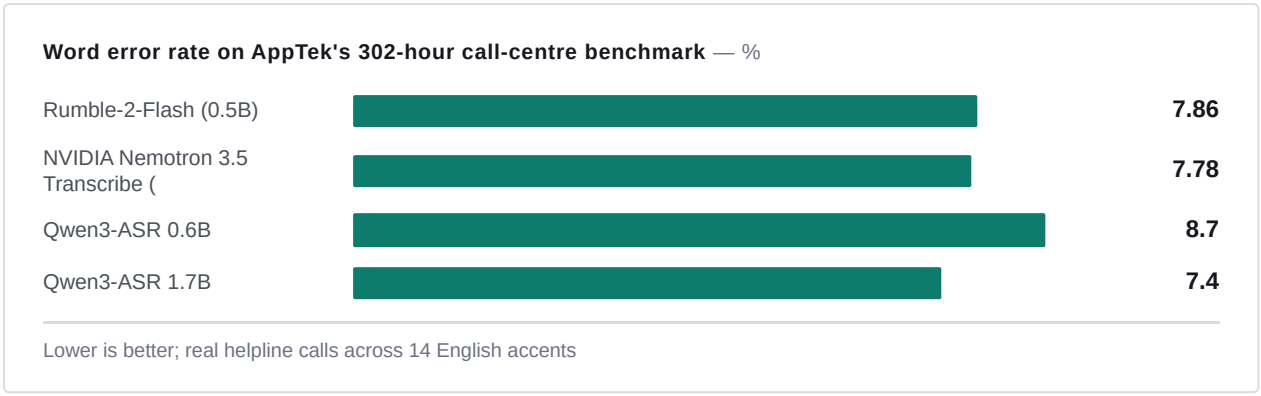
The pitch is cost. Transcription is one of the most common uses of AI in business — call centres, meeting notes, subtitles, voice assistants — and it is usually billed by the minute or run on expensive servers. A model this small that runs this fast changes that arithmetic: Rumble says it can process an hour of audio in about half a second when batches are full, which is the difference between transcribing calls overnight and transcribing them as they happen.

The most striking claim is about messy audio. On a 302-hour benchmark of real helpline calls, drawn from 14 English accents and scored by the benchmark's own script, Rumble-2-Flash had an error rate of 7.86%. NVIDIA's Nemotron 3.5 Transcribe, a 30-billion-parameter model, scored 7.78% — effectively the same. A 0.6-billion-parameter model from the Qwen3-ASR family scored 8.7%, and only a larger sibling at 1.7 billion did better, at 7.4%. Clean studio audio is easy for modern systems; compressed, noisy phone calls with varied accents are where they usually fall apart, and that is the case Rumble is making.

On the Open ASR Leaderboard's multilingual track, the model averages a 6.96% error rate across nine languages including English, Chinese, German, French, Spanish and Hindi — ahead of Whisper large-v3 and of Meta's omniASR-LLM 7B, which is 15 times larger. The English track averages 6.6%.

What is genuinely new is the method as much as the result. Rumble trained the small model by having it transcribe audio itself and letting its larger production model grade every word of that output, rather than

feeding it finished transcripts. That required a workaround, because the two models split text into different pieces; Rumble's fix, called GOLD, lets them be compared over the same text. The practical upshot for buyers is that a small, free model now sits within a fraction of a percentage point of much larger systems on the hardest everyday audio.



KEY FACTS

LICENCE
Apache 2.0, free for commercial use

LANGUAGES
40+, with automatic detection

SPEED
About half a second per hour of audio at full batches

CALL-CENTRE ERROR RATE
7.86%, against 7.78% for NVIDIA's 30B model

MULTILINGUAL ERROR RATE
6.96% across nine languages

WHERE TO GET IT
Hugging Face (Reson8/Rumble-2-Flash) or Rumble's API

> **WHY IT MATTERS**
>
> Transcription is a per-minute cost for call centres, media companies and anyone building voice features, and a free model that runs this fast and holds up on bad phone audio undercuts paid speech APIs. Teams currently paying for transcription should benchmark it against their own recordings before renewing contracts.

Read the original at huggingface.co →

---

04 MEDIUM IMPACT Google

Google opens Genie 4's interactive worlds to Ultra subscribers

> Google has opened its Project Genie world-generation tool to eligible AI Ultra subscribers, adding a fourth-generation model that can be redirected with a typed prompt while you are inside the world.

The update, announced on Google's DeepMind blog, gives paying Ultra subscribers first access to Genie 4, described by the company as its most realistic and interactive world model yet. The headline change is mid-session prompting: type a new instruction while navigating a generated environment and the weather shifts, a character appears or the action turns, in real time. Genie 4 also adds native audio that adapts as a scene unfolds, from ambient sound to character dialogue.

Genie 3, the earlier model, stays in a model picker for users who want longer runs or worlds seeded from images. Its sessions now stretch to three minutes per generation, up from 60 seconds, and separate worlds can be chained together for longer journeys. New worlds can also be grounded in 360-degree panorama images, and environments can be changed with natural-language prompts during exploration.

A new agent lets users describe a world in conversation rather than a single prompt; it coordinates Google's multimodal models to produce initial visuals and a background score matched to the mood. Gemini models can also sit inside a session, watching the screen to suggest where to go next or trigger events. A community gallery collects curated and user-made worlds, including one that turns personal photos into explorable spaces, a prehistoric habitat built from panoramas, and a desk-top game of rock-paper-scissors refereed by Gemini. Two named collaborators, creative technologist Will (BLVCKL!GHT) and filmmaker Anand Gandhi, built gallery pieces with early access.

What is genuinely new is the combination of live editing and audio inside a generated world, plus the ability to publish and remix other people's creations. What remains research-grade: Google says camera latency, intricate physical interactions, audio precision and long-term scene permanence are still being worked on. All generated video and audio carries SynthID watermarking, and inputs and outputs pass through safety screening. Google AI Pro members can join a waitlist.

KEY FACTS

AVAILABLE TO
Eligible Google AI Ultra subscribers, starting today

WAITLIST
Google AI Pro members

GENIE 3 SESSION LENGTH
Up to 3 minutes per generation, up from 60 seconds

NEW IN GENIE 4
Mid-session prompt editing and real-time native audio

WATERMARKING
SynthID on all generated video and audio

> **WHY IT MATTERS**
>
> Interactive world models are moving from demo to subscription product, which puts a new kind of content-creation tool in front of paying customers and gives Google a feature to justify its top-tier AI plan. For anyone building games, training simulations or marketing experiences, it is an early look at a workflow that could replace some hand-built 3D work.

Read the original at Google →

SECTION 2 OF 3

AI Tools & Ecosystem

3 stories ● 3 medium

05 MEDIUM IMPACT huggingface.co

Hugging Face runs a giant open model on a 32GB Mac

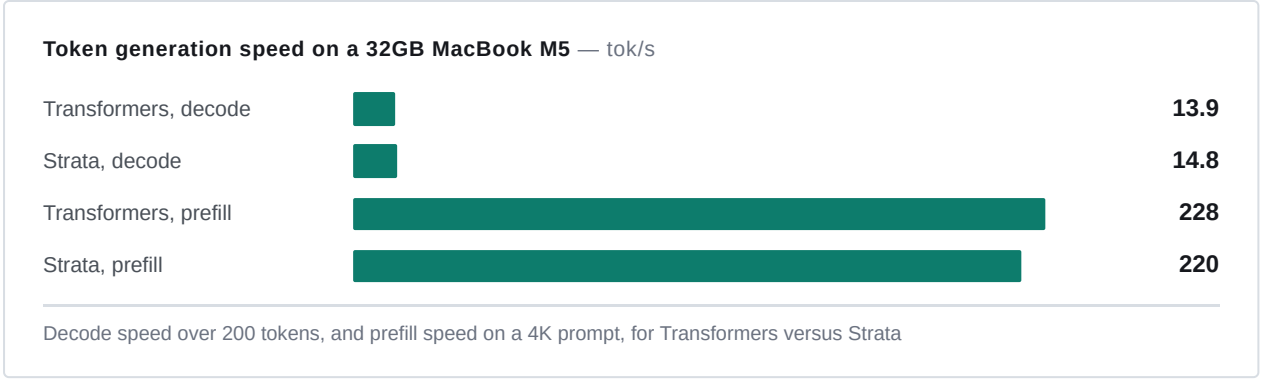
> A Hugging Face engineer has made a 180-billion-parameter model run on a 32GB MacBook by streaming most of its weights from the solid-state drive, coming within about 6% of a purpose-built inference engine on short replies.

The experiment, published on Hugging Face's blog, targets the kind of large 'mixture of experts' model — where each token is handled by a small handful of specialist sub-networks rather than the whole model — that normally needs far more memory than a laptop has. The author, Tarek Ziadé, modified the open-source Transformers library so that only the experts a given token actually uses are pulled from disk into a fixed-size cache, instead of copying every expert of every layer on every token, which is what the library did before. The change is a pull request, not a shipped release, and the API is still under discussion with maintainers.

The practical result: the model's experts, 121GB in their original form, were compressed to 3 bits and stored as 49GB on disk, with 12GB cached in memory. After loading, the setup used about 18GB of the Mac's 32GB, peaking at 25.4GB on a 6,000-token prompt — close enough to the limit that the author had to close his browser to stop macOS swapping. Speed reached 13.9 tokens per second on short replies, against 14.8 for Strata, a dedicated engine for this class of model; prefill, the stage where the model reads the prompt, was marginally faster than Strata at 228 versus 220 tokens per second.

The most surprising detail is how the gap was closed. Most of the remaining time was not spent fetching experts but waiting on the host: roughly 10 milliseconds per token lost while the graphics processor sat idle as Python queued the next layer's work. Forcing the GPU to commit work after each block added 10%, and custom kernels for two operations that lack a fast path on macOS helped further. The author says the real fix would need a native operation in PyTorch that it does not currently expose.

What is new here is the approach, not a product. Transformers could not do this out of the box; the project shows it can be done in Python with no C++ extension, and all the code was written with AI assistance. What is incremental is the payoff: at about 14 tokens per second, the author says it remains too slow for the coding agents he uses daily, and his two-GPU desktop still does the real work. The lesson for anyone weighing local models is that disk streaming makes larger models feasible on ordinary hardware, but not yet comfortable.



KEY FACTS

MODEL SIZE
180B parameters

HARDWARE USED
32GB MacBook M5

EXPERTS ON DISK AFTER 3-BIT COMPRESSION
49GB (from 121GB)

DECODE SPEED
13.9 tokens/sec vs 14.8 for Strata

GAP TO STRATA ON SHORT REPLIES
About 6%

CONVERSION TIME
About 5 hours

WHY IT MATTERS

It shows that the memory ceiling on running large open models locally is being pushed down toward ordinary laptops, which matters to anyone weighing cloud API bills against on-device inference — though at this speed it is a proof of concept, not a replacement for a workstation or a cloud service.

Read the original at huggingface.co →

---

06 MEDIUM IMPACT TechCrunch

Snapchat lets coding assistants build and publish camera effects

> Snapchat said on Wednesday that Lens creators can now build and publish augmented-reality effects using the AI coding assistants they already work in, including Claude Code, Cursor and Codex.

The tool, called Easy Lens CLI, extends Snapchat's existing Easy Lens, which turns a single written prompt into a Lens on the web or mobile. The new version puts that same process inside a developer's own coding setup: the assistant handles the technical work — assembling the effect from more than 100 AR building blocks and 50 AI asset generators, generating previews, testing, fixing errors and pushing updated versions to the creator's phone — while the creator describes what they want in plain language. It runs on Mac, Windows and Linux.

The pitch is reach. Snapchat's Lenses are the camera effects that let users change their appearance, drop digital objects into a room or play animated experiences, and building them has until now meant learning Snapchat's own development tools. By meeting developers inside the AI assistants they already use, Snapchat is trying to pull in people who would never have opened its toolchain — game developers porting existing browser games as Web Lenses, AI creators working from assets made in other tools, and agencies or brands turning campaign videos and images into AR with less production time. Snapchat says a game can go from idea to published Lens in a single day without the developer writing code.

The money detail is the one to repeat: Snapchat paid out $1.8 million to creators in its Lens+ Payouts program last month, and developers enrolled in that program can earn from Lenses made with Easy Lens CLI. That figure is small against the scale of most platform creator funds, which tells you Snapchat is still testing whether paid AR creation draws a durable community rather than competing on payouts.

What is genuinely new is the distribution channel, not the underlying idea of prompt-to-Lens. Snapchat is betting that the fastest way to grow its AR catalogue is to let outside AI tools do the building — a concession that its own developer tools were the bottleneck. For brands and agencies, the practical change is production time; for Snapchat, it is whether a wider pool of builders produces enough Lenses to keep the camera worth opening.

KEY FACTS

ANNOUNCED
Wednesday

TOOL
Easy Lens CLI

AI ASSISTANTS SUPPORTED
Claude Code, Cursor, Codex

PAID TO LENS+ CREATORS LAST MONTH
$1.8 million

PLATFORMS
Mac, Windows, Linux

BUILDING BLOCKS AVAILABLE
More than 100 AR blocks, 50 AI asset generators

> **WHY IT MATTERS**
>
> Brands, agencies and game developers can now turn existing assets into published Snapchat AR effects without learning Snapchat's tools or writing code, cutting production time. Snapchat is trading control of its developer workflow for a larger pool of Lens creators.

Read the original at TechCrunch →

---

07 MEDIUM IMPACT ollama.com

Ollama shows why small prompt changes inflate agent costs

> A timestamp or a mid-task settings tweak near the top of an AI coding agent's prompt can force the system to re-read the entire session from scratch, multiplying the computing work behind long agent runs, according to a joint post from Ollama and Together AI.

The two companies, which sit on opposite sides of the agent stack — Ollama decides what goes into each request, Together AI runs the inference engine that handles it — wrote up how small prompt changes break the cache that keeps long sessions affordable. Coding agents resend the whole conversation on every step: system prompt, tool definitions, files read, commands run. The reply is usually a short tool call. Caching the earlier work across turns is what stops cost and delay from climbing with every step.

The arithmetic in the post is the detail worth repeating. If each turn adds the same amount, ten turns without caching process 55 units of prompt; with caching, 10. At a hundred turns, which the authors say is not unusual for a coding task, it is 5,050 versus 100. The catch is that a cache matches from the start of the prompt up to the first token that differs and stops there. Change anything earlier, even one character, and everything after it is reprocessed. A current timestamp or git status in the system prompt re-rendered each turn means every turn is a full re-read.

The same trap applies to reasoning effort. In one model the setting is written into the first system block, so raising effort from low to high mid-task invalidates the whole session — hundreds of thousands of tokens spent on changing one setting. The authors advise picking an effort level at the start, or starting a fresh session at the higher setting. They also flag silent truncation of tool output: an agent handed the first 100 lines of a 640-line file without being told will confidently edit code it never saw. The fix is a line saying what was cut and how to request more.

Much of this is known to engineers who build agents and inference systems. What is new is the two sides putting it in one place and showing that neither can fix it alone: a harness can keep its prompt stable and still lose the cache if the engine routes turn seven to the wrong machine, and an engine can route perfectly and lose it if the harness changes the first line every turn. The post also notes a single GPU costs around $50,000 a year to rent, which is why idle time and wasted reprocessing carry a price.

KEY FACTS

PUBLISHED
November 16, 2026

AUTHORS
Parth Sareen and Zain Hasan, Together AI

PROMPT UNITS PROCESSED, 10 TURNS
55 without caching vs 10 with

PROMPT UNITS PROCESSED, 100 TURNS
5,050 without caching vs 100 with

ANNUAL RENT FOR ONE GPU
About $50,000

EXAMPLE FILE IN POST
640 lines, of which the agent saw 100

> **WHY IT MATTERS**
>
> Teams running coding agents at scale can cut both latency and GPU spend by keeping the start of the prompt fixed and avoiding mid-session setting changes; the same mistakes silently raise bills for anyone paying per token.

Read the original at ollama.com →

SECTION 3 OF 3

AI Applications & Industry

5 stories · 3 high · 2 medium

08 HIGH IMPACT TechCrunch

Nadella says Microsoft’s 60,000 buyout offers reflect an AI work redesign

> Microsoft offered buyouts to 60,000 employees in October, and chief executive Satya Nadella said Wednesday the aim was to let people choose whether to stay for a redesign of their jobs around AI.

Speaking at the DealBook Summit with Andrew Ross Sorkin, Nadella said the buyout was "to make sure that at least everyone felt like they had agency in making choices." He rejected the idea that Microsoft is swapping people for AI, software engineers especially. The company, he said, is reorganizing so that every employee uses AI to do the work and the job becomes managing that AI.

His example is the team behind Cowork, an AI agent inside Microsoft 365 apps. It has 11 developers who, in his words, "build a software factory that builds the product. They don't build the product." He expects that pattern to spread: "That level of direction is what, by the way, is going to happen in all jobs." For software engineering he described a shift in thinking — from disbelief that AI could write software, to alarm that it could do all of it, to a view that the discipline stays and thrives with a new "operating point" in how teams are organized and what engineers do.

The constraint is age. Microsoft is 51 years old and did not grow up as an AI-first company. Nadella said transforming the work, the workflow and the work output, with the pay structure that goes with it, "is going to take a bit more time than anyone expects," and some people will decide they have other choices to make. On The Pragmatic Engineer podcast with Gergely Orosz last week, he said the industry's rhetoric had "created enormous anxiety in people sort of feeling that they don't have control of their own destiny," naming talk of the singularity — the hypothetical point where AI surpasses human intelligence — and predictions that AI takes all jobs.

What is new is the framing, not the numbers: a sitting chief executive describing a voluntary exit program as the humane face of job redesign, and saying the redesign will reach every role. For workers at large established companies, the message is that AI adoption is being written into job descriptions rather than announced as layoffs. For investors, it signals slower restructuring than the rhetoric suggests, with costs carried through attrition and buyouts. For other employers, Microsoft has supplied a template — and a warning about how the conversation itself affects staff.

KEY FACTS

BUYOUT OFFERS
60,000 employees

OFFERED
October

COWORK TEAM SIZE
11 developers

COMPANY AGE
51 years

ANNOUNCED AT
DealBook Summit

> WHY IT MATTERS
>
> If you work at a large company, expect your role to be redefined around directing AI tools rather than doing the task yourself, with voluntary exits offered as the alternative. Employers watching Microsoft get a template for restructuring without announcing layoffs.

Read the original at TechCrunch →

---

09 HIGH IMPACT TechCrunch

Citi makes customers liable for AI-agent losses, even after hacking

> Citibank has rewritten its customer agreement so that anyone who hands a third-party AI agent access to their bank account carries the loss if the agent misbehaves — or gets hacked.

The change, made this week to Citi's Online User Agreement, covers tools that customers authorise to read their accounts, analyse spending, give investment advice or move money and pay bills. A Citi spokesperson said the update was meant to "provide greater clarity" about the risks of sharing banking credentials with outside AI tools, and to make clear that customers who grant that access are responsible for what the agents do with it.

Citi is not first. JPMorgan and Bank of America, the country's two largest banks, announced similar policies last month, and PNC, the eighth largest, changed its terms the same day as Citi. Citizens emailed customers this week with a comparable warning. What sets Citi's version apart is the hacking clause: if an AI tool leaks the customer's credentials or is itself breached, the customer is still on the hook for whatever thieves do with the account. That reverses the assumption many people carry into online banking — that a fraud department will clean up after a break-in.

The banks' discomfort has two sources. One is operational: automated assistants inside a bank's systems can make mistakes or act on bad instructions, and the banks are saying they will not unwind those transactions. The other is competitive. AI tools that compare fees or flag low-interest deposit accounts threaten the margins banks earn on customers who leave money where it sits, and markets reacted nervously when that prospect surfaced. If agents become the main way people make financial decisions, banks want to own that doorway. Citi's terms note that agents Citi itself provides carry no such extra liability for the customer.

The most telling detail is what a cautious AI company did. Anthropic launched a personal finance tool this month and billed it explicitly as "read-only" — it "can't move, manage, or invest your money." For anyone building consumer finance agents, the message is that read-only access is the version banks will tolerate, and transaction rights now come with the customer holding the risk. The pattern is familiar from other industries: AI labs promise to transform how people deal with service providers, and the service providers push back.

KEY FACTS

BANK CHANGING TERMS
Citibank, third largest US commercial bank

EFFECTIVE
This week

ALSO COVERED
JPMorgan, Bank of America (last month); PNC (same day); Citizens (email this week)

HACKING LIABILITY
Customer responsible even if AI tool leaks credentials or is breached

ANTHROPIC'S FINANCE TOOL
Read-only; cannot move, manage or invest money

CITI'S OWN AGENTS
No additional customer liability

> WHY IT MATTERS
>
> If you or your customers connect an AI assistant to a bank or brokerage account, the loss from a bad trade, a mistaken payment or a breach now sits with the account holder, not the bank. Read-only tools avoid the exposure; anything that can move money does not.

Read the original at TechCrunch →

---

10 HIGH IMPACT TechCrunch

Anthropic reportedly prices IPO at $83 for a $1.5 trillion valuation

> Anthropic sold shares at $83 apiece in an initial public offering that raised $81.3 billion, the most ever for a listing, valuing the Claude maker at roughly $1.5 trillion.

The price came in above the $70-to-$78 range Anthropic had floated, and the company sold 980 million shares rather than the 825 million it first planned. That combination — a higher price and more stock — pushed proceeds past SpaceX's $75 billion record. Trading on Nasdaq under the ticker "ANT" begins Thursday.

The valuation is a step up from the $965 billion Anthropic reached in May, when it raised $65 billion from private investors, though short of the $2 trillion it had reportedly sought. The listing also lands after the departure of Nvidia chief executive Jensen Huang unsettled confidence in AI stocks, making the pricing a test of whether public investors will pay up for AI exposure even in a shaky stretch. A strong first day would set the tone for private AI companies still weighing a listing, including rival OpenAI, which the source says was last reported valued at $1.25 billion and has not announced IPO plans.

The windfall spreads widely. Amazon owns nearly 15% of Anthropic and Google almost 14%, far more than any venture firm; no VC investor holds above 5%. Backers with roughly 1% stakes reportedly include Sequoia, former Google chief executive Eric Schmidt and Singapore's sovereign wealth fund GIC. Iconiq, Menlo Ventures, Coatue and Sound Ventures also stand to make billions, as does Spark Capital, the first institutional investor and the only VC firm known to hold a board seat. Sequoia, Iconiq, Lightspeed, Accel and Altimeter are reportedly planning to buy billions of dollars of shares in the offering itself. A group including AMD, Broadcom, Google, Micron, Microsoft, Nvidia and TSMC Partners agreed to buy $26 billion of non-voting shares in a private placement alongside the IPO.

The detail a reader will repeat: each of Anthropic's seven co-founders, including chief executive Dario Amodei, sold nearly 3.4 million shares in the offering, about $282 million apiece, according to the S-1 filing.

KEY FACTS

IPO PRICE
$83 a share

AMOUNT RAISED
$81.3bn

VALUATION
About $1.5 trillion

SHARES SOLD
980 million

TICKER AND VENUE
ANT on Nasdaq, trading from Thursday

FOUNDER PROCEEDS
About $282m each for seven co-founders

> WHY IT MATTERS
>
> The first big AI listing gives public investors a direct way to own a frontier lab, and its trading debut will shape what private rivals such as OpenAI can ask for. For anyone whose business depends on Claude or competing models, the new shareholder base adds pressure to show revenue growth in public.

Read the original at TechCrunch →

---

11 MEDIUM IMPACT TechCrunch

Humyn Labs raises $47 million to record robot-training data

> Humyn Labs, a Bengaluru startup that pays people to be filmed doing ordinary work and household chores, has raised $47 million to supply the footage as training data to U.S. robotics and AI developers.

The Series A combines $26 million of equity led by Accel with $21 million of venture debt led by Lighthouse Canton. Prosus Ventures, BeeGlobal, two Bajaj entities, Dallas Venture Capital, Jump Trading and Trifecta Capital also took part. The company was founded in February by Manish Agarwal, former chief executive of the Indian gaming firm Nazara Technologies, and Ishank Gupta, a former AB InBev executive.

The business exists because robot makers can get plenty of text and images to train software, but far less footage of hands opening jars, folding laundry or sorting parts — the kind of movement a machine needs to copy. Humyn Labs builds its own two-camera rigs to record people at work, then uses software to tag hand movements and object interactions so the clips become usable training material. It gathers footage in more than 20 countries across Latin America and Asia, with Brazil, India and South Korea as its main hubs, drawing on a contributor network built by the founders' earlier venture, KGeN.

The numbers are unusually concrete for a company this young. Humyn Labs says it booked $10 million of revenue between April and October, holds another $20 million in signed contracts and is negotiating a further $20 million. It has 12 paying customers, described as frontier AI labs, robotics companies and world-model developers; about 90% of revenue comes from the U.S. Gupta declined to name them. The company has 51 full-time employees, mostly in Bengaluru, plus local partners who find and supervise collection sites.

That gap between paying workers daily or weekly and waiting 30 to 60 days for lab customers to pay explains the debt slice: borrowing is less dilutive than selling more shares, Agarwal said. The most striking detail is the method. Rather than staging activities, the company films people doing their usual jobs and chores — which is cheaper and more realistic, but has drawn questions about privacy and consent when the recording happens inside homes. Agarwal's answer is supervision: workers are paid to take part, but need monitoring for the right input. India has become a hub for this work, with Human Archive and BuildAI among the others tapping its

large workforce. What is new here is the scale of revenue already attached to it; the data-collection model itself is spreading fast.

KEY FACTS

FUNDING RAISED
$47m

EQUITY PORTION
$26m, led by Accel

VENTURE DEBT
$21m, led by Lighthouse Canton

REVENUE, APRIL TO OCTOBER
$10m

SIGNED CONTRACTS
A further $20m

PAYING CUSTOMERS
12

> **WHY IT MATTERS**
>
> Physical-world training data is becoming a paid export business for emerging-market workers, and U.S. robot and model developers are the buyers. For anyone building robots, a supplier with signed contracts and U.S. customers is now a real option; for workers, the terms of consent and supervision are the issue to watch.

Read the original at TechCrunch →

---

12 **MEDIUM IMPACT** MIT Technology Review

Osmo says its AI nose matches people at distinguishing smells

> A New Jersey startup called Osmo says its AI model can now detect and tell apart smells about as well as the average person, and design new scents about as well as a junior perfume designer.

Osmo, founded in 2023 by Alex Wiltschko out of his work at Google Brain, has built what it calls the first foundation model for smell. The company's claim of "human-level olfactory intelligence" rests on two comparisons: the model is roughly as good as an average person at detecting and discriminating between odors, and roughly as good as a junior perfume designer at creating them. Osmo says its technology is already used in the fragrance industry, and that last month its model outperformed in-house perfumers at one of the world's largest consumer goods companies. It has also made retail fragrances, including for Procter & Gamble.

The work is expensive and manual. Osmo has assembled more than 10 million human labels and 800 million "scent tokens" — descriptions of the exact molecules and concentrations behind a given odor — by running spectrometers and other chemical sensors around the clock and having about 70 trained labelers in the US and overseas sniff everything from fruit to organs and write down what they smell. Before Osmo, Wiltschko counts fewer than 200,000 human descriptions of smells in existing databases, and the fragrance industry guards its olfactory research closely. The company also runs a smell-printing robot the size of a school bus that mixes ingredients to order; in a demo in Lisbon, a prompt for a Portuguese custard tart scent aimed at men aged 18 to 23 produced three candidate profiles and the molecules to make them, with a sample ready in about a week.

The most surprising finding is how much smells share. Wiltschko says the scent tokens defining a newborn baby overlap heavily with those of a fresh plate of cooked spaghetti — specifically Barilla brand, he notes. Elsewhere, the Gates Foundation, which put in $5 million at launch, later gave a $3.5 million grant to find insect repellents; Osmo identified eight substances more repellent than DEET. StockX has worked with the company on spotting counterfeit sneakers by the chemistry of their materials.

Not everyone buys the claims. Smell scientist Avery Gilbert has argued on his Substack that some marketing claims are undocumented and that Osmo's patent filings still lean on human sensory panels and perfumers. Michael Schmuker, a former professor who has consulted for Osmo, calls the technology an achievement but says smell depends on context — the molecule in Camembert also appears in used nappies, and only context keeps them apart. Wiltschko is candid that medical diagnostics, the biggest prize, has a long history of failed attempts, and casts Osmo's work as building at the foothills.

KEY FACTS

FOUNDED
2023, spun out of Google Brain

TRAINING DATA
Over 10 million human labels and 800 million scent tokens

HUMAN SMELL DESCRIPTIONS BEFORE OSMO
Fewer than 200,000, by founder's count

DATA LABELERS
Around 70 in the US and overseas

GATES FOUNDATION FUNDING
$5 million at launch, plus a later $3.5 million grant

INSECT REPELLENTS FOUND
Eight substances more repellent than DEET

> **WHY IT MATTERS**
>
> Fragrance and consumer-goods companies now have a tool that can generate and screen scent formulas faster than human perfumers, which cuts the cost and time of developing a new product. The bigger markets Wiltschko points to — airport threat detection and diagnosing disease by smell — remain unproven and depend on better sensors.

Read the original at MIT Technology Review →

Sources

01 Google’s Gemini 4 Neon cuts task costs by 32%
https://deepmind.google/blog/introducing-gemini-4-neon/

02 Claude solves two open problems in black-hole physics
https://www.anthropic.com/research/claude-black-hole-physics

03 Rumble releases a compact speech model for 40+ languages
https://huggingface.co/blog/rumble-2-flash

04 Google opens Genie 4’s interactive worlds to Ultra subscribers
https://deepmind.google/blog/genie-4-is-now-available/

05 Hugging Face runs a giant open model on a 32GB Mac
https://huggingface.co/blog/moe-expert-streaming

06 Snapchat lets coding assistants build and publish camera effects
https://techcrunch.com/2026/11/18/snapchat-now-lets-developers-build-ar-lenses-with-ai-coding-assistants-like-claude-code-and-cursor/

07 Ollama shows why small prompt changes inflate agent costs
https://ollama.com/blog/open-models-long-horizon-agents

08 Nadella says Microsoft’s 60,000 buyout offers reflect an AI work redesign
https://techcrunch.com/2026/11/18/microsoft-ceo-satya-nadella-explains-why-he-asked-60k-employees-to-quit/

09 Citi makes customers liable for AI-agent losses, even after hacking
https://techcrunch.com/2026/11/18/more-banks-says-customers-using-ai-agents-are-on-their-own-even-if-they-get-hacked/

10 Anthropic reportedly prices IPO at $83 for a $1.5 trillion valuation
https://techcrunch.com/2026/11/18/anthropic-reportedly-prices-ipo-at-83-a-share-above-projected-range/

11 Humyn Labs raises $47 million to record robot-training data
https://techcrunch.com/2026/11/18/accel-backs-indias-humyn-labs-to-supply-training-data-needed-by-u-s-robotics-ai-developers/

12 Osmo says its AI nose matches people at distinguishing smells
https://www.technologyreview.com/2026/11/18/1148816/smell-startup-osmo-ai-nose-good-as-humans/

About this document. Every story in the 19 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 2026-11-19T06:37:47Z · model deepseek-v4.1-flash:cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.