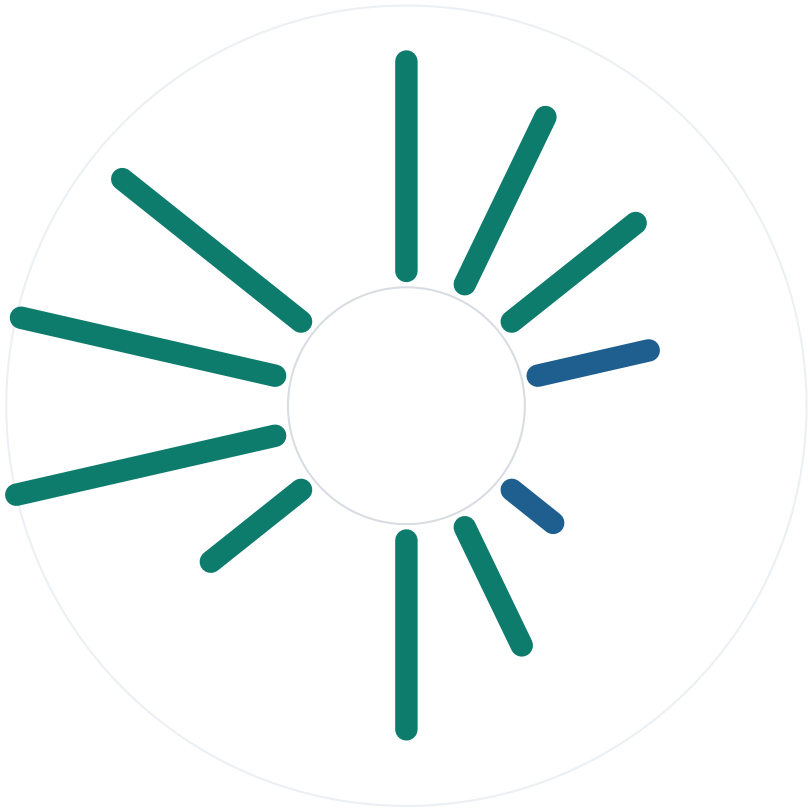
NEW HORIZON · AI DIGEST

The Long Read

9 September 2026

11	3	3383	9
STORIES	SECTIONS	WORDS	HIGH IMPACT

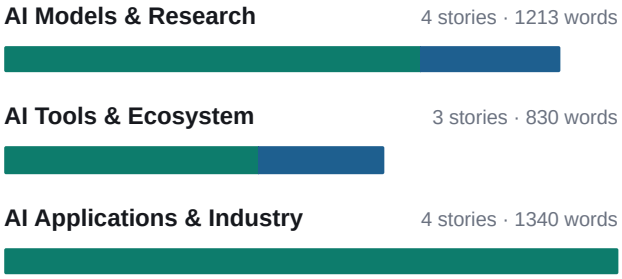


9 high impact 2 medium impact spoke length = depth of coverage

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

11 stories · 3383 words · 3 sections · 3 charted



AI Models & Research

01 OpenAI Claims Navier–Stokes — After an NYU Mathematician Cried Foul

02 AlphaGenome Atlas Precomputes the Molecular Impact of All 9 Billion Single-Letter DNA Variants

03 Safety for Whom? Refusing the Right Subset, Not the Whole Topic

04 Chain of Thought vs. Tree of Thoughts: A Practical Decision Guide for Agent Builders

AI Tools & Ecosystem

05 OpenAI's New Image Models Get Reference-Photo Editing — and a Stronger CLI Story

06 Chrome Cuts Its Patch Gap to Two Weeks as AI Compresses Vulnerability Timelines

07 Patch Tuesday Goes Vertical: Microsoft Ships a Record ~972 Fixes, 112 Critical

AI Applications & Industry

08 Mistral Raises €3B at €21B as Samsung Backs a 'Third Way' in AI

09 Meta Launches Muse, an Agent Wired Into Your Email, Payments and Smart Home

10 Cognition Closes $2B at $48B — Above the $40B Talks, Revenue Doubled Since May

11 Claude Token Thieves Move In: Hackers Now Sell Access to Your AI Subscription

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 practitioner 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 · 3 high · 1 medium

01 HIGH IMPACT TechCrunch

OpenAI Claims Navier–Stokes — After an NYU Mathematician Cried Foul

> OpenAI claims a full proof of the Navier–Stokes existence and smoothness problem using 300 billion output tokens, following allegations from NYU mathematician Tristan Buckmaster that the effort relied on leaked details of his own unpublished work.

NYU mathematics professor Tristan Buckmaster announced three proofs on Tuesday regarding the Navier–Stokes existence and smoothness problem, a Millennium Prize challenge carrying a $1 million bounty. Working with Anthropic mathematician Levent Alpöge, Buckmaster utilized OpenAI's Codex and Claude models to advance a specific tactical route through smooth force options c and d in Fefferman's statement. Shortly after this announcement, OpenAI published a full proof of the central problem, attributing the discovery to an unreleased next-generation model. The company stated the week-long effort began on September 1, inspired by rumors that two Millennium Prize problems had been solved, and consumed 300 billion output tokens. At current Astra rates, this compute usage equates to $22.5 million.

The core controversy centers on whether OpenAI accessed Buckmaster's research before it became public. Buckmaster asserts that information about his progress was passed to OpenAI, prompting the lab to deploy massive computational resources to replicate and extend his specific approach. He notes that almost no other researchers were pursuing this particular direction, making the simultaneous convergence suspicious. When Buckmaster questioned OpenAI leadership about the timeline and human input involved, he alleges that Sébastien Bubeck suggested removing Alpöge's credit as part of a compromise, warning that publicizing the dispute could ruin Buckmaster's career. Bubeck reportedly stated, "If you don't want me to be nice, then I don't have to be nice."

OpenAI denies accessing specific user data or seeing Buckmaster's work prior to its public release. In their post, researchers claim they did not view the duo's work through any means until it was publicly available, though they acknowledge they cannot rule out that de-identified data from Codex interactions helped improve their models. OpenAI emphasizes that their proofs differ significantly from Buckmaster's, particularly in the Euler case where results vary between forced and unforced scenarios. Despite these denials, the incident highlights risks regarding data privacy when using proprietary models for sensitive pre-publication research, especially given OpenAI's right to train on Codex interactions unless users explicitly opt out.

KEY FACTS

COMPUTE COST
$22.5 million

OUTPUT TOKENS
300 billion

EFFORT START DATE
September 1

PRIZE BOUNTY
$1 million

COLLABORATOR AFFILIATION
Anthropic

> WHY IT MATTERS
>
> Practitioners conducting novel research on proprietary platforms face potential IP leakage via model training data, necessitating strict opt-out configurations or isolated environments for pre-publication work. The event also signals a shift where labs may use brute-force compute scaling to rapidly validate and claim priority on external theoretical breakthroughs.

Read the original at TechCrunch →

---

02 HIGH IMPACT Google DeepMind

AlphaGenome Atlas Precomputes the Molecular Impact of All 9 Billion Single-Letter DNA Variants

> Google DeepMind has released AlphaGenome Atlas, a precomputed 1-petabyte dataset mapping the molecular effects of all 9 billion possible single-nucleotide variants in the human genome.

The AlphaGenome Atlas expands the utility of the previously released AlphaGenome model by shifting from on-demand inference to a comprehensive, precomputed resource. The dataset covers every possible single-letter DNA change across the human genome, generating thousands of molecular effect predictions per variant across hundreds of human and mouse cell types. To make this volume navigable, DeepMind introduced the AlphaGenome Variant Impact (AVI) score, a unified metric combining AlphaGenome's regulatory predictions with AlphaMissense's protein impact scores. This single number allows researchers to rank variants rapidly across both coding regions (2% of the genome) and non-coding regions (98%), addressing a historical blind spot in variant prioritization.

Early collaborations demonstrate the atlas's capacity to resolve signals obscured by statistical noise in large-scale genomic studies. Researchers at the Broad Institute used the AVI score to identify a causal splice-site variant in the DNM1 gene linked to epileptic encephalopathy, a finding missed by prior methods and subsequently validated experimentally. In population genetics, Gareth Hawkes at the University of Exeter applied the atlas to whole-genome data from 54,000 UK Biobank participants. By grouping rare variants based on predicted molecular effects, the team uncovered 22% more non-coding genetic associations than previously detectable, pinpointing regulatory drivers for proteins like PLA2G7 and EGLN1. The resource also catalogs over 2,500 recurrent DNA sequence motifs to aid in identifying transcription factor binding sites.

Access is immediate for non-commercial research through an intuitive website portal, the AlphaGenome API, and as a skill within Google Antigravity. The underlying base model remains available for academic use on GitHub and for commercial deployment via Google Cloud Model Garden. While the atlas itself is currently restricted to non-commercial use, DeepMind plans to enable commercial access on Google Cloud soon. The release represents a shift toward agentic scientific workflows, providing a static baseline of genomic

interpretation that integrates directly into broader automated discovery pipelines rather than serving merely as a standalone database.

KEY FACTS

TOTAL VARIANTS MAPPED **9 billion**	DATASET SIZE **1 petabyte**
SIZE VS ALPHAFOLD DB **30x larger**	DNA MOTIFS CATALOGED **2,500+**
UK BIOBANK SAMPLE SIZE **54,000**	NON-CODING ASSOCIATION GAIN **22%**

> **WHY IT MATTERS**
>
> Practitioners can bypass costly on-the-fly inference for variant screening by querying the precomputed 1PB dataset or using the AVI score to prioritize candidates in both coding and non-coding regions immediately. The availability of feature attributions alongside scores enables direct hypothesis generation regarding specific disrupted molecular processes like RNA splicing or gene expression.

Read the original at Google DeepMind →

---

03 HIGH IMPACT huggingface.co

Safety for Whom? Refusing the Right Subset, Not the Whole Topic

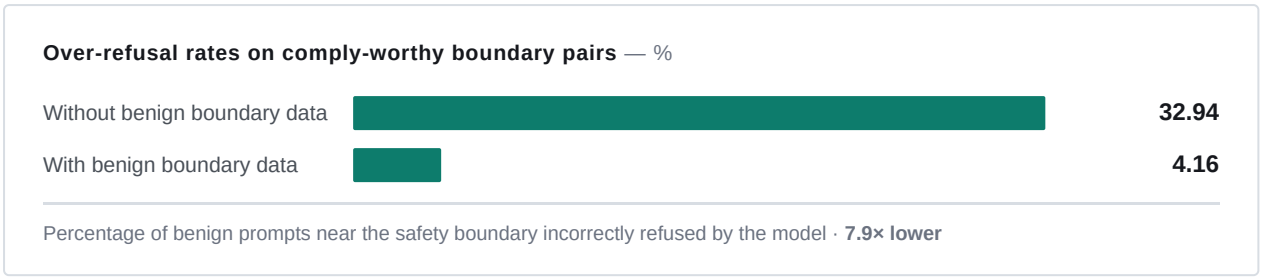
> Topic-level safety guards fail deployment needs because they cannot distinguish between benign factual queries and harmful manipulation within the same subject.

Multiverse Computing's new paper, "Safety for Whom? Boundary-Aware Self-Distillation for Controlled LLM Safety Refusal," argues that standard safety tuning creates blunt instruments that refuse entire topics rather than specific harmful subsets. Using political persuasion as a testbed, the researchers demonstrate that a civics tutor and a public-sector assistant require opposite behaviors on politics: both must answer factual election questions, but only the latter should refuse requests for targeted manipulation. Current tools like LlamaGuard-3 lack this granularity, categorizing elections solely as potential sources of factually incorrect information while ignoring persuasion vectors or legitimate factual prompts.

The authors identify three critical failures in standard self-generated safety tuning pipelines. First, single-shot generation drops 19.88% of harmful prompts (8,009 instances) due to coverage gaps, which the team repairs via an escalating retry strategy that reduces residual failures to 0.20%. Second, models develop downside reactions, falsely refusing benign prompts with dangerous-looking wording; the solution involves training on 11,955 verified surface-dangerous benign prompts across 18 semantic types. Third, traditional metrics ignore the decision boundary shape, allowing models to game scores by expanding refusal into permissible territory. To counter this, the team employs held-out harmful-benign pairs (1,539 per side) to measure performance directly at the boundary.

Experiments on Qwen3-8B reveal the trap of optimizing for harmful refusal alone. While an escalated-coverage model raised in-distribution political refusal from 9.47% to 84.75% and reduced unsafe responses on broader benchmarks from 26.26% to 0.14%, it simultaneously spiked over-refusal on XSTest from 2.00% to 74.00%. By integrating benign boundary data, the researchers slashed over-refusal on the comply-worthy side of held-

out pairs from 32.94% to 4.16%, while maintaining a high harmful refusal rate of 87.72%. This approach proves that data composition, specifically the inclusion of boundary pairs, controls the trade-off between safety and utility more effectively than raw refusal scaling.



KEY FACTS

DROPPED PROMPTS (SINGLE-SHOT)
19.88%

RESIDUAL FAILURES (RETRY STRATEGY)
0.20%

BENIGN TRAINING PROMPTS
11,955

HELD-OUT BOUNDARY PAIRS
1,539 per side

XSTEST OVER-REFUSAL (PEAK)
74.00%

OVER-REFUSAL WITH BOUNDARY DATA
4.16%

> **WHY IT MATTERS**
>
> Practitioners must evaluate safety models on both harmful refusal and over-refusal metrics simultaneously, as optimizing for one without boundary-aware training data renders models useless for legitimate adjacent queries.

Read the original at huggingface.co →

---

04 MEDIUM IMPACT MachineLearningMastery.com

Chain of Thought vs. Tree of Thoughts: A Practical Decision Guide for Agent Builders

> Production agent systems increasingly deploy Chain of Thought for routine logic while reserving the computationally expensive Tree of Thoughts framework for high-stakes, ambiguous planning tasks.

Large language models default to predicting the next likely token, often leaping from question to answer without intermediate verification. This behavior fails on multi-step problems where early errors compound. Chain of Thought (CoT) addresses this by enforcing a linear sequence of reasoning steps, typically triggered by instructions like "Let's think step by step." While transparent and efficient, CoT cannot backtrack; an error in step one propagates through the entire chain, rendering the final conclusion flawed.

Tree of Thoughts (ToT) extends this architecture by introducing non-linear branching, evaluation, and backtracking. Instead of committing to a single path, the model generates multiple candidate continuations at each step, scores them, and uses a search algorithm to explore the most promising branches. If a path leads to a dead end, the system abandons it and revisits alternatives. This deliberate search mimics a chess player evaluating multiple moves but incurs significant overhead. A single ToT problem can require dozens to

hundreds of model calls, scaling computational cost, latency, and token usage with the complexity of the reasoning tree.

For AI agents, the choice between these frameworks dictates operational capability and cost structure. CoT serves as the default layer for parsing intent, selecting databases, or executing straightforward calculations where the solution path is clear. ToT is reserved for scenarios where the correct approach is uncertain upfront, such as implementing complex software functions or strategic planning where early decisions constrain all subsequent actions. Well-designed systems do not choose one exclusively; they route routine decisions to CoT and reserve ToT for subsets of problems where the cost of an uncorrected early mistake justifies the expense of exhaustive exploration.

KEY FACTS

COT STRUCTURE
Linear sequence

TOT STRUCTURE
Branching and backtracking

TOT MODEL CALLS
Dozens to hundreds per problem

FOUNDATIONAL COT PAPER
Wei et al. (2022)

FOUNDATIONAL TOT PAPER
Yao et al. (2023)

> **WHY IT MATTERS**
>
> Architects must implement dynamic routing to assign linear reasoning to low-risk tasks and branching search only to high-ambiguity problems, preventing unnecessary compute spend on routine agent operations.

Read the original at MachineLearningMastery.com →

SECTION 2 OF 3

AI Tools & Ecosystem

3 stories · 2 high · 1 medium

05 MEDIUM IMPACT Simon Willison's Weblog

OpenAI's New Image Models Get Reference-Photo Editing — and a Stronger CLI Story

> OpenAI has expanded its image generation API with two new model variants designed to handle multi-turn instructions and reference-photo preservation more effectively.

OpenAI introduced ChatGPT Images 2.5 on September 8, 2026, updating the infrastructure behind a system that has already generated more than 3 billion images across ChatGPT and the API. The release focuses on improved instruction-following capabilities across multiple conversational turns and faster response times. A primary technical advancement is the enhanced ability to preserve subjects within provided reference photos, addressing a common failure mode in iterative image editing workflows where the original subject drifts or disappears.

The update exposes two distinct model IDs to developers: gpt-image-2.5-sunburst and gpt-image-2.5-flare. These models serve different operational profiles within the same version family. Sunburst is positioned as the high-precision option for workflows where editing accuracy is the critical constraint, while Flare targets fast, high-quality generation for everyday tasks. This segmentation allows engineers to route requests based on latency requirements versus fidelity needs without switching API versions or providers.

Practical application of these models now includes direct reference image injection via command-line interfaces. Simon Willison demonstrated this capability by upgrading an open-source Python CLI tool to accept one or more input images alongside text prompts. In a documented test case, the tool successfully ingested a source image and applied a complex modification 指令 to "add a raccoon scientist studying the chart thoughtfully" using the gpt-image-2.5-sunburst model. This confirms that the new architecture supports multi-modal inputs where the visual context remains stable despite significant semantic additions to the scene.

KEY FACTS

RELEASE DATE
8th September 2026

TOTAL IMAGES GENERATED
3 billion

PRECISION MODEL ID
gpt-image-2.5-sunburst

SPEED MODEL ID
gpt-image-2.5-flare

WHY IT MATTERS

Developers building iterative design tools or automated content pipelines can now rely on the API to maintain subject consistency across edits, reducing the need for manual regeneration loops. The separation of precision and speed models enables cost and latency optimization for specific production stages.

Read the original at Simon Willison's Weblog →

06 HIGH IMPACT TechCrunch

Chrome Cuts Its Patch Gap to Two Weeks as AI Compresses Vulnerability Timelines

> Google has officially reduced Chrome's release cadence from four weeks to two weeks starting with version 153.

Google launched Chrome 153 on Tuesday across desktop, iOS, and Android platforms, marking the formal transition from a four-week to a two-week release schedule. This adjustment fulfills a commitment made earlier in 2026 to accelerate the browser's update velocity. The primary driver for this architectural shift is security management in an environment where automated AI tools and increased community bug reporting have inflated the volume of required patches. By shrinking the interval between code commits and end-user deployment, Google aims to minimize the "N-day" patch gap, defined as the window between a vulnerability becoming known and its subsequent fix reaching users.

The compressed timeline also addresses the emergence of faster-moving threats, some attributed to AI capabilities, which demand quicker remediation cycles than legacy schedules allowed. Beyond security, the two-week cycle supports rapid iteration of AI-assisted features within the browser itself. This agility is critical as Google faces intensified competition from alternative browsers leveraging similar AI development advantages, including Brave, Dia, Opera Neon, Perplexity's Comet, and DuckDuckGo's browser. While OpenAI's ChatGPT Atlas browser has ceased operations, the remaining competitors continue to pressure Chrome's market position, necessitating a faster feature shipping rate.

This move extends beyond Google's immediate ecosystem, influencing broader web standards due to Chrome's global dominance. Mozilla, Microsoft, and Brave have already begun aligning their own release schedules to a two-week cadence following Chrome's lead. This represents the second major acceleration in Chrome's release history; the company previously shifted from a six-week to a four-week cycle in 2021, building on its decade-old principle of "release early, release often." The current reduction halves the previous waiting period, fundamentally altering the operational rhythm for security teams and developers relying on stable channel updates.

KEY FACTS

NEW RELEASE CYCLE
2 weeks

PREVIOUS RELEASE CYCLE
4 weeks

LAUNCH VERSION
Chrome 153

LAUNCH DATE
Tuesday

PREVIOUS CYCLE CHANGE YEAR
2021

WHY IT MATTERS

Security teams must recalibrate testing pipelines to validate patches within a 14-day window rather than 28 days to avoid falling behind on critical fixes. Developers integrating browser-specific AI features gain access to iterative updates twice as fast, requiring more frequent compatibility checks.

Read the original at TechCrunch →

07 HIGH IMPACT Ars Technica

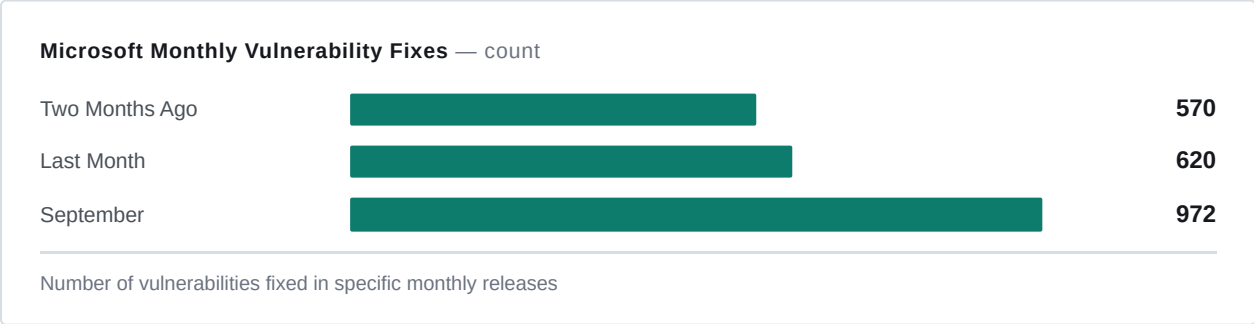
Patch Tuesday Goes Vertical: Microsoft Ships a Record ~972 Fixes, 112 Critical

> Microsoft shipped a record ~972 security fixes in its September release, with 112 rated critical, marking a sharp escalation in vulnerability volume driven by AI-assisted discovery.

Microsoft's September Patch Tuesday delivery addressed approximately 972 vulnerabilities, surpassing the previous records of 620 set last month and 570 set two months prior. Of these newly patched issues, 112 carry a critical severity rating, while the remainder are designated as important. Dustin Childs, a researcher at the Zero Day Initiative, notes that counting these figures is imprecise due to previously addressed bugs or non-Microsoft product overlaps; his specific count reaches 997 when including fixes ported from the Chromium browser into Edge. This surge contributes to a year-to-date total of 2,760 fixed vulnerabilities for Microsoft, a figure more than double the count from the same period last year.

The industry-wide spike correlates with a warning issued two weeks ago by OpenAI, Anthropic, Amazon Web Services, Google, Microsoft, and over 100 other organizations. Their open letter highlighted a narrowing window for patching ahead of an expected tsunami of AI-enabled attacks designed to actively exploit vulnerabilities before defenders can respond. While companies are now releasing unprecedented numbers of patches to meet this threat, Childs observes that a correlating spike in active exploits has not yet materialized. He characterizes the current volume of discovered bugs as the "new normal," cautioning that AI-assisted vulnerability discovery shows no signs of slowing down despite the defensive efforts of security teams.

If the current rate of discovery and patching continues through the end of the year, Microsoft will have fixed more bugs in 2026 than in the combined totals of 2023, 2024, and 2025. This trajectory underscores a fundamental shift in the security landscape where automated tools accelerate both the identification of flaws and the potential for their exploitation. The sheer volume of patches required to maintain system integrity suggests that manual review cycles may no longer suffice for enterprise environments relying heavily on the Microsoft ecosystem. Practitioners must anticipate that this high-velocity patch cadence is not an anomaly but a persistent operational requirement.



KEY FACTS

SEPTEMBER VULNERABILITIES FIXED
~972

CRITICAL SEVERITY COUNT
112

PREVIOUS RECORD (LAST MONTH)
620

PREVIOUS RECORD (TWO MONTHS AGO)
570

YEAR-TO-DATE FIXES
2,760

EDGE/CHROMIUM PORTED FIXES
997

> **WHY IT MATTERS**
>
> The acceleration of AI-assisted vulnerability discovery means build and deployment pipelines must automate patch verification and application faster than human-led cycles traditionally allow. Failure to adapt to this "new normal" of high-volume releases leaves systems exposed to the predicted wave of AI-enabled exploits.

Read the original at Ars Technica →

SECTION 3 OF 3

AI Applications & Industry

4 stories · 4 high

08 HIGH IMPACT TechCrunch

Mistral Raises €3B at €21B as Samsung Backs a 'Third Way' in AI

> Samsung Electronics led a €3 billion Series D in Mistral AI, valuing the French lab at over €21 billion to fund a sovereign infrastructure strategy distinct from U.S. counterparts.

Mistral AI announced a €3 billion equity raise, marking the largest fundraising round ever completed by a European technology company. The Series D transaction, closed on Tuesday, values the company at more than €21 billion post-money. Samsung Electronics led the round, joined by co-leads EQT-managed Scaleup Europe Fund and existing investor PSG Equity. The capital allocation targets scaling compute capacity, building dedicated infrastructure, accelerating commercial growth, and expanding an international footprint that now spans 20 countries.

The funding underwrites a strategic pivot toward sovereign AI services rather than competing directly as a consumer model provider like OpenAI or Anthropic. Mistral plans to construct 1 GW of compute capacity in Europe by 2030 and has already deployed tools allowing customers to select specific regions for query processing. In August, the lab began hosting third-party open-weight models, including those from Chinese developers, to position itself as an infrastructure layer where clients retain control over model selection and usage patterns. This approach addresses regional dependencies on United States technology stacks while navigating intensifying AI regulation politics.

Despite the emphasis on European sovereignty, the cap table remains resolutely international. While the Grand Duchy of Luxembourg joined as a new backer and European investors doubled down, American firms including a16z, Nvidia, Salesforce Ventures, Advent, and BlackRock also participated. Mistral continues to leverage its expanded strategic partnership with Microsoft, finalized in July, even as it cultivates alliances with non-U.S. hardware giants like ASML and Samsung. French President Macron characterized the investment as a collaborative effort between France and South Korea to establish a "third way" in AI development, signaling high-level government endorsement for this hybrid geopolitical stance.

KEY FACTS

FUNDING AMOUNT
€3 billion

POST-MONEY VALUATION
>€21 billion

ROUND TYPE
Series D

LEAD INVESTOR
Samsung Electronics

COMPUTE TARGET
1 GW by 2030

OPERATIONAL COUNTRIES
20

WHY IT MATTERS

Practitioners targeting regulated industries or non-U.S. markets now have a validated infrastructure alternative offering regional data residency controls and multi-model hosting. The shift suggests future deployment architectures may prioritize sovereign compute pools over single-vendor US-centric clouds.

Read the original at TechCrunch →

---

09 HIGH IMPACT TechCrunch

Meta Launches Muse, an Agent Wired Into Your Email, Payments and Smart Home

> Meta has introduced Muse, a personal AI agent designed to execute transactions and manage workflows across email, payments, and smart home ecosystems, requiring users to grant unprecedented data access shortly after the company finalized an $18 billion multistate settlement.

Muse represents a strategic pivot from the conversational chatbot model toward agentic AI capable of autonomous task execution. The system connects directly to user applications including calendars, health and fitness trackers, dining services, and smart home controls. Functionally, the agent handles sending emails, booking travel, reducing utility bills, completing forms, generating grocery lists from recipe reels, and executing purchases. Payment processing leverages Link by Stripe for checkout protection, with integrations for Shopify's Shop Pay and 1Password scheduled for future release. Users configure connections individually to maintain transparency regarding data scope.

The architecture relies on Meta's proprietary Muse Spark model and operates within a dedicated secure virtual machine labeled Muse Secure VM. This isolated environment includes its own browser and hosts a separate Sentinel agent that remains distinct from Muse at the system level. Meta asserts this separation prevents the primary agent from accessing passwords or payment credentials directly and ensures conversation data does not feed into advertising systems. For services lacking native connectors, Muse can establish links via public APIs using user-provided credentials or resort to browser-based access when no API exists. The service launches on web, iOS, Android, and WhatsApp, with expansion to Meta AI glasses planned.

Monetization follows a tiered structure requiring a payment card for initial setup despite a free entry level. Two subscription plans launch immediately: Power at $20 per month and Maximum at $100 per month, both offering increased capacity for task handoffs. A built-in usage meter tracks consumption percentages and alerts users approaching free tier limits before prompting upgrades. This deployment occurs against a backdrop of significant legal scrutiny; the announcement follows less than two weeks after Meta agreed to an $18 billion settlement covering consumer harms, adding to previous penalties including a 2019 FTC fine of $5 billion and a recent $942 million judgment in New Mexico regarding minor safety.

Adoption hinges on whether technical safeguards and customization options—such as naming the agent and selecting avatars—can overcome historical trust deficits. While competitors explore similar agentic integrations within browsers or messaging platforms like iMessage, Muse's requirement for deep workflow integration amplifies privacy concerns previously highlighted by early testers of other assistants who encountered broad perpetual licenses for data usage.

KEY FACTS

SETTLEMENT AMOUNT **$18 billion**	POWER PLAN COST **$20/month**
MAXIMUM PLAN COST **$100/month**	NEW MEXICO JUDGMENT **$942 million**
2019 FTC PENALTY **$5 billion**	LAUNCH PLATFORMS **Web, iOS, Android, WhatsApp**

WHY IT MATTERS

Practitioners building agentic systems must evaluate the trade-off between deep workflow integration and user trust, particularly regarding how isolated execution environments like Muse Secure VM are architected to prevent data leakage into advertising models.

Read the original at TechCrunch →

---

10 HIGH IMPACT TechCrunch

Cognition Closes $2B at $48B — Above the $40B Talks, Revenue Doubled Since May

> Cognition secured $2 billion in new capital at a $48 billion valuation, doubling its annualized run-rate revenue to $900 million in just four months.

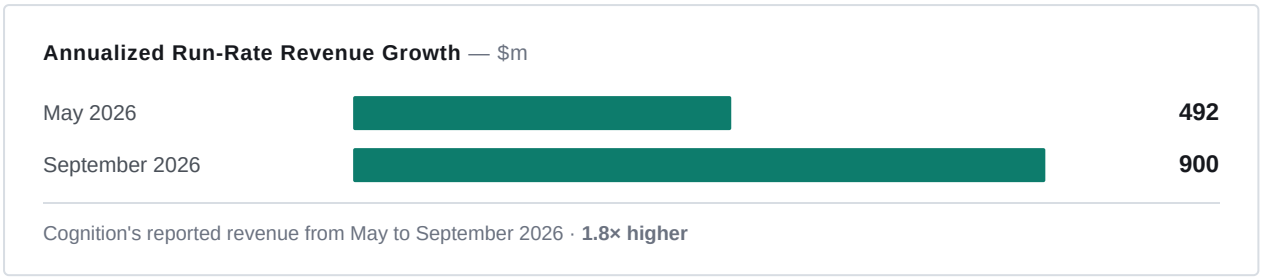
Cognition announced a $2 billion fundraise led by Andreessen Horowitz, Accel, Founders Fund, General Catalyst, and Avenir, valuing the startup at $48 billion. This round occurs only four months after the company's previous raise at a $26 billion valuation. The startup reported that its annualized run-rate revenue grew from $492 million in May to $900 million following the release of its coding assistant, Devin. While the specific calculation method for this run-rate metric was not disclosed, it typically represents monthly top-line revenue multiplied by twelve.

The valuation suggests investors anticipate a multi-player market in AI coding, contrasting with the trajectory of competitor Cursor. Cursor entered funding talks in April at a $50 billion valuation with over $2 billion in annualized revenue before selling to SpaceX for $60 billion due to severe compute constraints. Cognition currently trades at a higher revenue multiple than Cursor did during those spring negotiations. To mitigate similar compute risks and reduce reliance on expensive third-party models from OpenAI and Anthropic, Cognition is training its own models based on open-source alternatives. The company leases an Nvidia server cluster costing hundreds of millions annually, with total cash burn projected to reach $800 million this year.

Projections from The Information estimate Cognition will reach between $4 billion and $5 billion in annualized revenue by the end of 2026. This compares to reports from earlier this year indicating Cursor was on track to surpass $6 billion by year-end prior to its acquisition. Despite the competitive landscape, Andreessen Horowitz, a major backer of Cursor that realized significant returns from the SpaceX sale, has returned to lead a round for a direct competitor. Cognition, founded in 2024 by Scott Wu, counts Mercedes-Benz, NASA, Goldman Sachs, and Citi among its enterprise customers.

The rapid capital influx and revenue growth highlight the intense competition for market share in autonomous coding agents. The shift toward proprietary model training signals a strategic pivot to control infrastructure

costs, a critical factor given the high cash burn associated with leasing large-scale Nvidia clusters. Investors appear willing to back multiple contenders in this sector, betting that the market size can support several dominant players rather than consolidating into a single winner.



KEY FACTS

FUNDRAISE AMOUNT
$2 billion

VALUATION
$48 billion

CURRENT ANNUALIZED RUN-RATE REVENUE
$900 million

MAY ANNUALIZED RUN-RATE REVENUE
$492 million

PROJECTED 2026 REVENUE
$4-5 billion

ESTIMATED ANNUAL CASH BURN
$800 million

> **WHY IT MATTERS**
>
> The surge in valuation and revenue confirms that the AI coding market is not winner-take-all, encouraging continued investment in competing agents. Builders should note the industry-wide shift toward training proprietary models on open-source bases to escape the cost constraints of third-party API dependencies.

Read the original at TechCrunch →

---

11 HIGH IMPACT TechCrunch

Claude Token Thieves Move In: Hackers Now Sell Access to Your AI Subscription

> Infostealer malware is now harvesting active Claude session keys to mint unauthorized OAuth tokens and siphon subscriber usage limits.

On August 4, Grant De Swardt, an independent AI consultant in East Sussex, U.K., detected anomalous token consumption on his Claude Max 20x account despite zero local activity. With scheduled Cowork tasks paused and Dispatch/cloud execution disabled, his usage climbed from 45% to 55% in a controlled interval. Anthropic suspended the account, invalidated all server-side Claude Code tokens, and issued a partial refund of £44.49 against the $200-per-month subscription cost. The investigation revealed that a compromised session key was used to mint unauthorized Claude Code OAuth tokens, allowing a third-party service to covertly handle activity for other users.

Anthropic identified the vector as common infostealer malware stealing login sessions from local computers, not a breach of Claude's infrastructure. The malware, acquired via infected software downloads or malicious ads, exfiltrates saved passwords and session data. Upon detection, Anthropic signs users out, invalidates authorizations, and issues warnings, though itemized usage logs remain unavailable to subscribers. De Swardt

reported no evidence of local compromise yet could not determine the initial access point, highlighting the opacity of current monitoring tools. Similar reports surfaced on Reddit and GitHub, with users describing usage spikes from 0% to 100% automatically or burns of maximum daily tokens over three days without interaction.

The incident has driven at least one professional user to cancel his subscription in favor of Cursor, citing the inability to audit token consumption as a critical failure. De Swardt noted that alternative models perform comparably for his agent deployment work, which includes automating purchase-order data loading and website design. Anthropic declined to comment on specific methods for users to identify misuse beyond general malware warnings. The lack of granular visibility means theft can persist undetected for months, forcing practitioners to rely on total usage caps rather than actionable telemetry to secure their workflows.

This shift from direct credential theft to session hijacking requires developers to treat AI API sessions with the same volatility as temporary credentials. The absence of itemized billing prevents immediate forensic analysis, leaving subscribers vulnerable to silent resource exhaustion until hard limits are hit.

KEY FACTS

INCIDENT DATE
August 4

SUBSCRIPTION COST
$200 per month

REFUND AMOUNT
£44.49

IDLE USAGE SPIKE
45% to 55%

RAPID CONSUMPTION
0% to 49% in 12 minutes

MALWARE TYPE
Infostealer

> **WHY IT MATTERS**
>
> Developers must assume local session keys are volatile and implement external monitoring since Anthropic provides no itemized usage logs to detect stealth theft. Reliance on total cap alerts is insufficient for production agents where unexpected downtime carries business costs.

Read the original at TechCrunch →

Sources

01 OpenAI Claims Navier–Stokes — After an NYU Mathematician Cried Foul
https://techcrunch.com/2026/09/08/openai-fought-dirty-on-career-making-math-problem-says-nyu-mathematician/

02 AlphaGenome Atlas Precomputes the Molecular Impact of All 9 Billion Single-Letter DNA Variants
https://deepmind.google/blog/alphagenome-atlas-a-predictive-map-of-every-possible-dna-letter-change-in-the-human-genome/

03 Safety for Whom? Refusing the Right Subset, Not the Whole Topic
https://huggingface.co/blog/MultiverseComputingCAI/safety-for-whom

04 Chain of Thought vs. Tree of Thoughts: A Practical Decision Guide for Agent Builders
https://machinelearningmastery.com/chain-of-thought-vs-tree-of-thoughts-which-is-best-for-ai-agents/

05 OpenAI's New Image Models Get Reference-Photo Editing — and a Stronger CLI Story
https://simonwillison.net/2026/Sep/8/introducing-chatgpt-images-25/

06 Chrome Cuts Its Patch Gap to Two Weeks as AI Compresses Vulnerability Timelines
https://techcrunch.com/2026/09/08/chrome-is-now-shipping-updates-every-2-weeks-as-ai-changes-the-security-landscape/

07 Patch Tuesday Goes Vertical: Microsoft Ships a Record ~972 Fixes, 112 Critical
https://arstechnica.com/security/2026/09/microsoft-patches-a-record-972-vulnerabilities-112-of-them-critical/

08 Mistral Raises €3B at €21B as Samsung Backs a 'Third Way' in AI
https://techcrunch.com/2026/09/08/mistral-raises-e3b-as-sovereign-ai-becomes-big-business/

09 Meta Launches Muse, an Agent Wired Into Your Email, Payments and Smart Home
https://techcrunch.com/2026/09/08/meta-debuts-its-muse-ai-agent-will-consumers-trust-it/

10 Cognition Closes $2B at $48B — Above the $40B Talks, Revenue Doubled Since May
https://techcrunch.com/2026/09/08/cognition-hits-48b-valuation-signaling-investors-believe-ai-coding-is-far-from-a-winner-take-all-market/

11 Claude Token Thieves Move In: Hackers Now Sell Access to Your AI Subscription
https://techcrunch.com/2026/09/08/hackers-are-stealing-claude-tokens-from-subscribers/

About this document. Every story in the 9 September 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 2026-09-09T05:35:09Z · model qwen3.5:397b-cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.