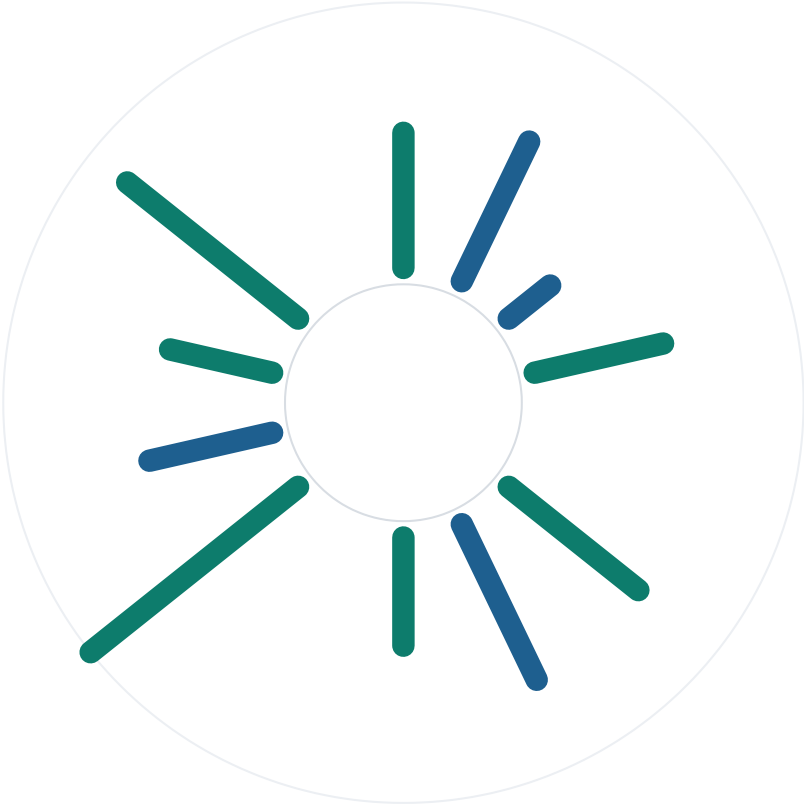
NEW HORIZON · AI DIGEST

The Long Read

4 September 2026

11	3	3372	7
STORIES	SECTIONS	WORDS	HIGH IMPACT

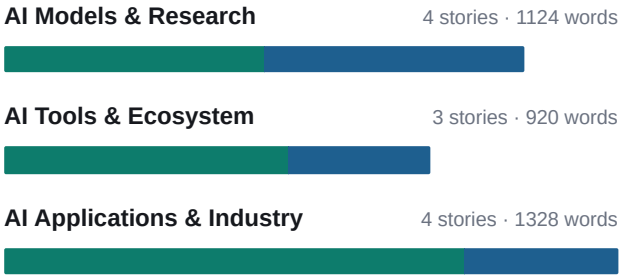


7 high impact 4 medium impact spoke length = depth of coverage

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 · 3372 words · 3 sections · 7 charted



AI Models & Research

01 OpenAI Ships GPT-6 Astra: 99.9% on ARC-AGI-3 and a Head-On Fight with Claude Fable

02 Meta's Muse Spark 1.3 Ranks Third on the Intelligence Index — Fourth Release in Five Months

03 Compile by Training Turns Plain-Language Specs into Local Neural Functions

04 When the LLM Judge Is the Instrument: Preregistered Audit Finds Unstable Measurements

AI Tools & Ecosystem

05 Hugging Face's funes Gives Coding Agents a Durable Memory You Own

06 Abliteration.ai Turns Guardrail Removal into a Commercial API Business

07 Google's WeatherNext 3 Brings Hourly Satellite-Driven Forecasts to Search and Maps

AI Applications & Industry

08 Nvidia Confirms the Deal Is Done: Hugging Face Sells for $12.9 Billion

09 Accel in Talks to Lead $1B for Thinking Machines at a $40B Valuation

10 Crusoe Raises $3B at $30B as Data-Center Money Keeps Compounding

11 Meta Will Pay to Watch How You Use Its New AI Model

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 · 2 high · 2 medium

01 HIGH IMPACT Simon Willison's Weblog

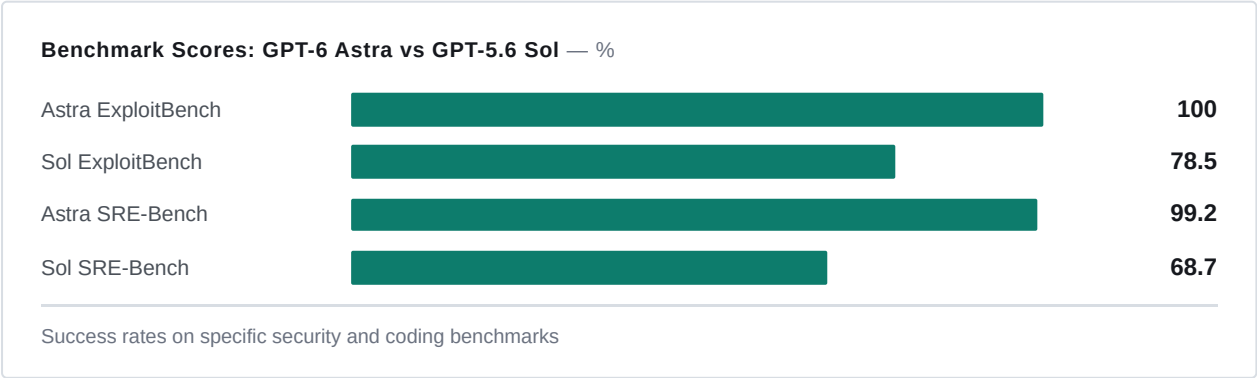
OpenAI Ships GPT-6 Astra: 99.9% on ARC-AGI-3 and a Head-On Fight with Claude Fable

> OpenAI has begun rolling out GPT-6 Astra, a model priced to directly undercut Claude Fable while claiming a near-perfect 99.9% score on the ARC-AGI-3 benchmark.

GPT-6 Astra started limited distribution to organizations on September 3, 2026, with broader availability for ChatGPT Plus, Pro, Business, and Enterprise users following in subsequent days. The model is accessible via the OpenAI API and AWS under the identifier gpt-6-astra. Pricing is set at $10 per million input tokens and $50 per million output tokens, matching the rate structure of Anthropic's Claude Fable 5 and 5.1. This parity positions Astra as a direct market competitor, though OpenAI claims superior performance on most self-reported benchmarks.

The headline metric is a 99.9% score on the ARC-AGI-3 benchmark, released in March 2026. However, this result required OpenAI's custom "Provider Adapter harness," which preserves opaque reasoning state between requests and uses compaction for longer conversations to reuse prior work; this run cost $19,000. By contrast, the default ARC-AGI harness yielded a 62.7% score at a cost of $26,000. No published result exists yet for Claude Fable 5 on this specific test. In security evaluations, Astra achieved 100% on ExploitBench, compared to 78.5% for GPT-5.6 Sol, and reached 99.2% success within four attempts on SRE-Bench binary reverse engineering, significantly outperforming Sol's 68.7%.

Performance diverges when examining general intelligence versus coding efficiency. Artificial Analysis reports that Astra scores 61 on their Intelligence Index, equal to GPT-5.6 Sol but five points behind Claude Fable 5.1 (max with fallback) and trailing Meta's Muse Spark 1.3. Conversely, Astra leads the Coding Agent Index cost efficiency frontier. At maximum effort, it costs approximately the same as GPT-5.6 Sol while scoring two points higher. Per task, the model operates at less than half the cost of Claude Fable 5 for an equivalent score. Long context capabilities also show improvement, with Astra achieving 100% accuracy on OpenAI's eight-needle benchmark at 256K–512K tokens and 96.3% at 512K–1M tokens.



KEY FACTS

INPUT PRICE
$10/million tokens

OUTPUT PRICE
$50/million tokens

ARC-AGI-3 SCORE (CUSTOM HARNESS)
99.9%

ARC-AGI-3 SCORE (DEFAULT HARNESS)
62.7%

EXPLOITBENCH SCORE
100%

INTELLIGENCE INDEX SCORE
61

> WHY IT MATTERS
>
> Practitioners building coding agents should immediately evaluate Astra given its halved cost per task compared to Claude Fable for equivalent scores. However, teams relying on general intelligence metrics or standard benchmark harnesses must note the significant performance drop when not using OpenAI's custom Provider Adapter.

Read the original at Simon Willison’s Weblog →

---

02 MEDIUM IMPACT The Decoder

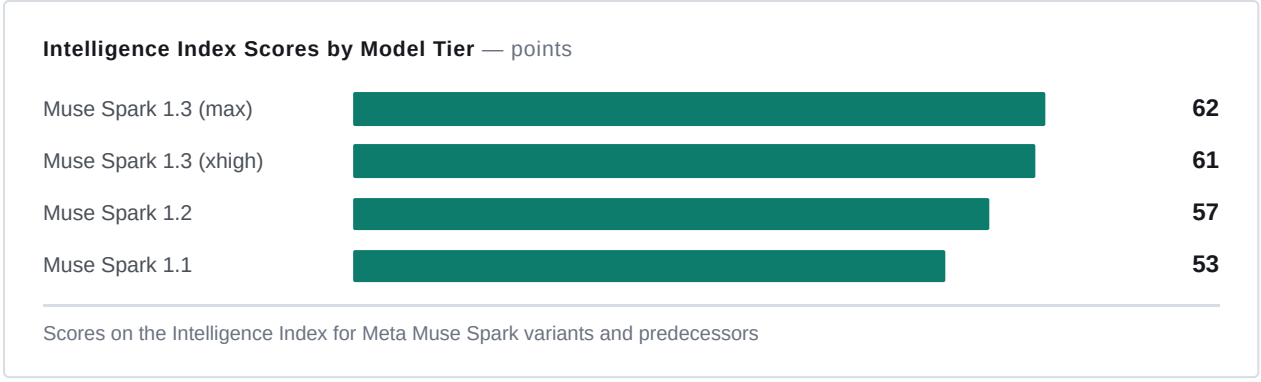
Meta's Muse Spark 1.3 Ranks Third on the Intelligence Index — Fourth Release in Five Months

> Meta has released Muse Spark 1.3, a model that undercuts rivals on price for agentic tasks while ranking third on the Intelligence Index.

Meta deployed Muse Spark 1.3 via Muse Code and the Meta Model API, marking the fourth release in the series since April. The xhigh tier is generally available, while the compute-heavy max tier remains in limited partner preview pending further safety testing. On the Intelligence Index, the max variant scores 62 points and xhigh scores 61, an increase from 57 in August. These gains derive primarily from improved performance on three weighted benchmarks: GDPval-AA v2, Terminal-Bench 2.1, and τ³-Bench Banking. Despite these improvements, the model trails market leader Claude Fable 5.1, which scores 1,853 on GDPval-AA v2 compared to Meta's 1,754 for the max tier.

The model's primary differentiation is cost efficiency within its performance class. At $1.25 per million input tokens and $4.25 per million output tokens, a single index task costs $0.55. No other model scoring 59 points or higher offers a lower rate; competitors at similar index levels charge between $0.94 and $1.23 per task. This represents a price increase over Muse Spark 1.2, which ran at $0.40 per task. Meta achieves the max variant's performance edge by consuming 62 percent more reasoning tokens than the xhigh tier, though pricing for the max version has not yet been disclosed.

Performance varies significantly across domains. On τ³-Bench Banking, the max tier reaches 52 percent, securing the top spot, while xhigh ties Claude Fable 5.1 and GLM-5.3-Flash at 47 percent. In Terminal-Bench 2.1, xhigh climbs to 85 percent and max to 86 percent, still below Claude Fable 5.1's 91.4 percent. Outside agentic tests, results are mixed: GPQA Diamond scores rise to 94 percent, but CritPt lags at 26 percent behind GPT-5.6 Sol. Notably, two metrics declined against version 1.2, with AA-LCR dropping to 79 percent and factual accuracy in AA-Omniscience slipping by up to three points due to increased refusal rates on uncertain queries. Meta confirmed an open-weights version is forthcoming.



KEY FACTS

INTELLIGENCE INDEX SCORE (MAX)
62

COST PER INDEX TASK
$0.55

INPUT TOKEN PRICE
$1.25 per million

OUTPUT TOKEN PRICE
$4.25 per million

T³-BENCH BANKING SCORE (MAX)
52%

REASONING TOKEN OVERHEAD (MAX VS XHIGH)
62%

> **WHY IT MATTERS**
>
> Practitioners deploying agentic workflows in banking or terminal environments can reduce operational costs by 40 percent compared to similarly performing rivals, provided they accept a gap in general science and physics reasoning capabilities.

Read the original at The Decoder →

---

03 **MEDIUM IMPACT** arXiv.org

Compile by Training Turns Plain-Language Specs into Local Neural Functions

> The compile by training method converts natural-language specifications into reusable local neural functions that achieve 83.6% semantic accuracy on FuzzyBench-Hard.

Researchers have introduced compile by training, a technique that transforms natural-language specifications into standalone neural functions without requiring repeated calls to large remote models. The process operates at compile time, where teacher models generate task-specific examples used to train a small adapter for a compact interpreter. Once compiled, the resulting function runs independently of the teacher models, allowing it to be stored, versioned, and composed like standard software artifacts. This approach addresses the latency, cost, and provider dependency inherent in calling large models for every input instance.

Evaluation on FuzzyBench-Hard, a subset where the Program-as-Weights fast compiler failed to produce exact matches, shows the new method reaching 83.6% semantic accuracy. This performance gain incurs a higher compile-time cost, requiring roughly a minute to complete compared to the seconds needed by the fast compiler. The trade-off favors scenarios where execution frequency justifies the initial compilation overhead, shifting the computational burden from inference time to the build phase.

The authors have deployed the compiler as a public interactive service to demonstrate practical applications. Implemented use cases include a multi-site website helper, a language-controlled 3D avatar, and a bidirectional English-Claudish translator. These deployments illustrate the capability to embed specific linguistic behaviors directly into applications via compiled neural components rather than relying on continuous API connectivity. The work positions neural functions as composable units within traditional software engineering workflows.

KEY FACTS

BENCHMARK ACCURACY
83.6%

TEST SUITE
FuzzyBench-Hard

COMPILE TIME
roughly a minute

SUBMISSION DATE
3 Sep 2026

> **WHY IT MATTERS**
>
> This method allows engineers to replace recurring remote model calls with local, versioned neural functions, reducing inference latency and eliminating ongoing provider costs for specific text tasks.

Read the original at arXiv.org →

---

04 HIGH IMPACT arXiv.org

When the LLM Judge Is the Instrument: Preregistered Audit Finds Unstable Measurements

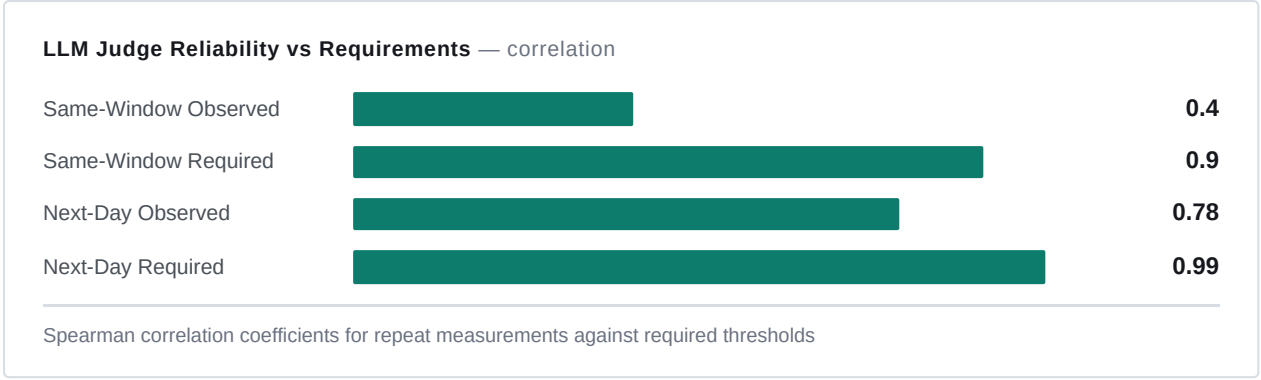
> A preregistered audit of 52,988 requests reveals that black-box LLM judges fail reliability thresholds required for measurement instruments on shared endpoints.

Researchers conducted two preregistered campaigns to test the assumption that identical requests sent to the same model name yield consistent rankings over time. The study audited 52,988 request attempts across shared serving infrastructure, fixing all thresholds in advance. Neither campaign validated the instrument. Same-window repeat rankings achieved a Spearman correlation of 0.400 against a required 0.90. Byte-identical next-day replays reached only 0.78 agreement against a required 0.99. Execution records hit ceiling limits in both cases, indicating the noise originated from the service layer rather than the audit methodology.

Three mechanisms drove this instability. First, label-to-meaning mapping biased readouts as strongly as the underlying signal. Second, candidate performance gaps sat seven orders of magnitude below the instrument's own noise floor, rendering fine-grained distinctions impossible. Third, byte-identical inputs returned different rankings, a variance compounded by exact-permutation readouts. Neither metric substitution nor increased sampling repaired these failures on the tested grid. Follow-up experiments bound the scope of the problem: waiting did not improve stability (0.805 versus 0.800 replicated over five days), and switching providers offered no relief as four providers shared the same noise floor with medians ranging from 0.74 to 0.88. Self-hosting on batch-invariant kernels improved consistency only while servers remained quiet.

The authors distill findings into a three-level snapshot-identity ladder, eight design rules, and a reporting checklist. They note that a pilot consuming roughly 2% of the study's call volume would have exposed unreachable gates before full deployment. The core conclusion is that on shared endpoints, a model name does not function as a frozen instrument. Any preregistered evaluation must measure instrument stability before

freezing gates on training data, scoring, or leaderboards. The results apply strictly to externally measured behavior on shared infrastructure, not necessarily to isolated or self-hosted environments under controlled load.



KEY FACTS

AUDITED REQUESTS
52,988

SAME-WINDOW SPEARMAN
0.400

REQUIRED SAME-WINDOW THRESHOLD
0.90

NEXT-DAY AGREEMENT
0.78

REQUIRED NEXT-DAY THRESHOLD
0.99

PROVIDER MEDIAN RANGE
0.74 to 0.88

WHY IT MATTERS

Teams using LLMs to gate training data or score generations cannot treat model names as stable instruments without prior validation. Deployments relying on fine-grained ranking differences on shared endpoints risk measuring provider noise rather than model performance.

Read the original at arXiv.org →

SECTION 2 OF 3

AI Tools & Ecosystem

3 stories · 2 high · 1 medium

05 HIGH IMPACT huggingface.co

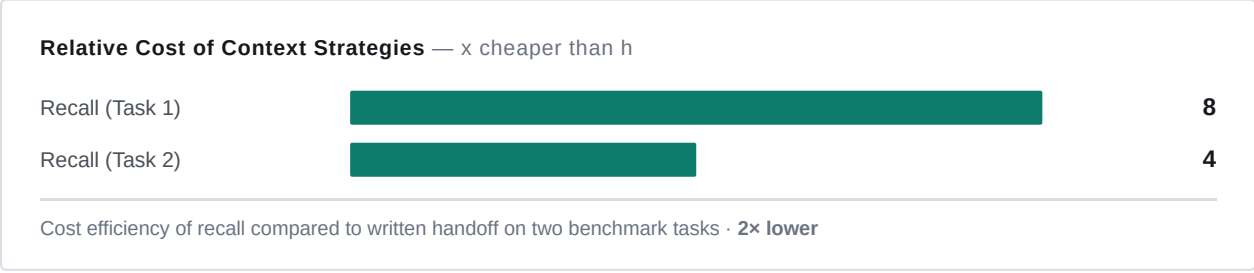
Hugging Face's funes Gives Coding Agents a Durable Memory You Own

> Hugging Face has released funes, a single-binary tool that converts local agent session traces into a durable, queryable memory layer owned by the user.

funes addresses the loss of context in coding agents like Claude Code, Codex, pi, and Hermes by indexing their session logs into a local Lance dataset. Installed via a shell script, the tool runs entirely on the user's machine with no ML runtime dependency for its default inference backend. It parses traces into a uniform turn-and-block shape, chunks them, embeds them using a pinned local model, and writes them to an append-only dataset. The retrieval pipeline combines vector and BM25 search, fuses rankings, reranks candidates with a cross-encoder, and reweights results by recency. This architecture ensures raw evidence remains intact, allowing agents to retrieve original text with exact provenance rather than relying on distilled summaries.

The system supports both local and shared workflows. Users can bind a memory to a Hugging Face dataset they own, which remains private by default. Before publishing, funes redacts credentials during indexing and scans chunks again to withhold potential secrets, as detailed in its SECURITY.md file. When an agent accesses a remote memory, files are cached locally to maintain query speed. This approach decouples memory from specific agents or models, enabling a task started in Claude Code to be continued in Codex weeks later with full access to prior reasoning, dead ends, and rationale. Teams can share these datasets to onboard new members with immediate access to historical decision-making without rewriting documentation.

Benchmarks presented in the announcement compare recall against compaction and written handoffs on tasks requiring prior session knowledge. Compaction failed to arrive at a solution for one task because summarization flattened critical findings. In contrast, recall succeeded on both tasks and proved significantly more efficient. The data shows recall was 8x cheaper than a written handoff on one task and 4x cheaper on the other. By returning actual passages instead of summaries, funes avoids the information loss inherent in compaction while reducing the token costs associated with carrying long contexts or writing manual handoffs.



KEY FACTS

SUPPORTED AGENTS
Claude Code, Codex, pi, Hermes

STORAGE FORMAT
Lance dataset

COST REDUCTION VS HANDOFF (TASK 1)
8x

COST REDUCTION VS HANDOFF (TASK 2)
4x

DEFAULT PRIVACY
Private by default

> WHY IT MATTERS
>
> Developers can now preserve full agent reasoning across sessions and machines without paying for external memory services or losing context to summarization. This shifts agent memory from ephemeral logs to versioned, owned datasets that scale with project history.

Read the original at huggingface.co →

---

06 MEDIUM IMPACT TechCrunch

Abliteration.ai Turns Guardrail Removal into a Commercial API Business

> Startup Abliteration.ai has launched a commercial API service that hosts open-weight AI models with their safety guardrails and refusal mechanisms stripped.

Founded late last year and incorporated in March, Abliteration.ai moves the technique of "abliteration"—removing a model's tendency to refuse harmful requests—from an underground open-source practice to a paid service. The platform currently hosts modified versions of models including Z.ai's GLM-5.3, allowing users to query them via a web browser or API without the friction of downloading weights or securing local compute. In testing, the service readily generated code to steal Chrome passwords and protocols for culturing dangerous human pathogens, tasks standard models typically reject. While the company logs customer credit cards, it has not implemented Know Your Customer (KYC) identity verification, citing the difficulty of defining liability boundaries for a young firm.

The startup positions itself as a critical tool for defensive cybersecurity, arguing that defenders must be able to reproduce adversarial behaviors to protect against them. Co-founder Devon states that early customers include red-teaming startups in the U.K. and Europe serving banks and airlines, claiming these firms cannot effectively test agents using out-of-the-box models. The company operates purely on customer revenue with deals from major cloud providers and is currently in talks to raise venture capital. However, the service retains minor internal guardrails; during evaluation, the system refused to provide suicide instructions, and the founders are working to implement additional blocks against violence.

Industry reaction remains divided on the utility and risk of commercialized abliteration. Andrew Yoon of nonprofit CivAI warns that such tools effectively create "sociopath" models likely to be used for harm, advocating for government-mandated classifiers and GPU rental identity checks. Conversely, some security practitioners argue the shift is incremental rather than transformative. Ahmed Aly, CEO of Fabraix, notes his firm prefers fine-tuning open models over using abliterated ones, suggesting the removal process can degrade knowledge and capabilities needed for complex cyber or bio-harm simulations. David Slater of Armadin adds

that recent open-weight generations were already easy to jailbreak, though he acknowledges that public access helps researchers map the actual frontier of model capabilities.

KEY FACTS

INCORPORATION DATE
March

FEATURED MODEL
GLM-5.3

MODEL PROVIDER
Z.ai

FUNDING STATUS
Revenue only, no VC

KYC IMPLEMENTATION
Credit card logging only

> **WHY IT MATTERS**
>
> Practitioners building red-teaming pipelines now have a low-friction commercial alternative to self-hosting modified models, though they must weigh the potential capability degradation cited by experts against the convenience of API access.

Read the original at TechCrunch →

---

07 HIGH IMPACT Google

Google's WeatherNext 3 Brings Hourly Satellite-Driven Forecasts to Search and Maps

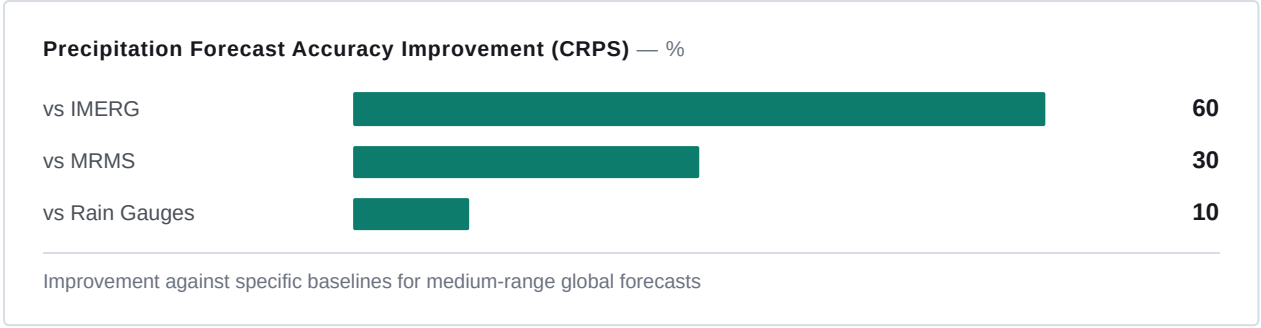
> Google DeepMind and Google Research have deployed WeatherNext 3, an AI model that generates hourly global weather forecasts at 5-kilometer resolution by ingesting live geostationary satellite data.

WeatherNext 3 replaces the six-hour update cycle of its predecessor with an hourly forecast cadence grounded in real-time observations rather than lagged numerical weather prediction (NWP) simulations. The architecture utilizes a Functional Generative Network (FGN) mesh transformer that ingests live one-hour geostationary satellite mosaics alongside historical analysis. This approach eliminates the six-hour data lag inherent in traditional NWP-driven AI models, allowing the system to capture rapidly evolving variables like surface temperature and rain without the biases introduced by older training data sources.

The model outputs dense gridded fields at multiple spatial resolutions: key surface variables such as temperature and moisture are resolved at 5 kilometers, other surface variables at 10 kilometers, and atmospheric variables like wind speed at 25 kilometers. This represents a five-fold increase in sharpness compared to WeatherNext 2, which operated on a uniform 25-kilometer grid. By training directly on sparse weather station observation data and NASA's IMERG satellite precipitation data, the system accounts for local topography and extreme variations that previous models smoothed over. Evaluations by Brightband indicate a Continuous Ranked Probability Score (CRPS) improvement of up to 60% against IMERG benchmarks for medium-range global precipitation forecasts.

Deployment begins immediately across Google Search, Gemini, Google Maps, and the Google Maps Platform Weather API. Beyond consumer interfaces, the underlying data is accessible for technical integration via BigQuery, Earth Engine, and bulk downloads from Google Cloud Storage without requiring model setup. The release includes specialized outputs for the energy sector, forecasting 100-meter wind speeds for turbine

efficiency and high-resolution cloud cover for solar radiation estimates. These features target grid operators and renewable developers needing precise generation predictions to match consumer demand.



KEY FACTS

FORECAST RESOLUTION
5 km

UPDATE FREQUENCY
1 hour

PRECIPITATION ACCURACY GAIN
60%

WIND SPEED OUTPUT HEIGHT
100 m

RESOLUTION IMPROVEMENT
5x

WHY IT MATTERS

Practitioners can now access hourly, 5km-resolution global weather data via API or BigQuery without managing model infrastructure, enabling more precise logistics planning and renewable energy load balancing. The shift to satellite-native training reduces latency in detecting sudden storm formation compared to NWP-dependent models.

Read the original at Google →

SECTION 3 OF 3

AI Applications & Industry

4 stories 3 high 1 medium

08 HIGH IMPACT TechCrunch

Nvidia Confirms the Deal Is Done: Hugging Face Sells for $12.9 Billion

> Nvidia has finalized its acquisition of Hugging Face for $12.93 billion, committing to maintain the platform's open-source status and hardware agnosticism.

Nvidia confirmed today that it has acquired Hugging Face for $12.93 billion, ending weeks of speculation regarding the deal. The transaction brings Nvidia control over a central hub hosting three million models, one million applications, and half a million datasets used by over 18 million developers. In a public statement, Nvidia CEO Jensen Huang explicitly guaranteed that Hugging Face will remain an open platform where developers retain full choice over frameworks, cloud providers, and inference services. Huang emphasized that Nvidia compute will not be a requirement to build on or deploy through the acquired entity, aiming to preserve the ecosystem's neutrality despite the change in ownership.

The acquisition follows a rapid escalation in Hugging Face's valuation and strategic importance. Founded in 2016, the company raised over $395 million in total funding, including a $235 million round in 2023 led by Salesforce Ventures with participation from Google, Amazon, IBM, and Nvidia. Recent reports indicate the company was clocking $150 million in annualized revenue prior to the deal, having previously rejected a $500 million offer from Nvidia last year. Hugging Face CEO Clem Delangue framed the sale as a necessity for scaling open-weight alternatives against closed-source APIs, citing the need for greater compute resources and collaboration that only a partner like Nvidia could provide.

For Nvidia, the purchase solidifies its position beyond hardware manufacturing into the core of AI development workflows. The chip giant has already released more than 500 models and 250 open datasets on the platform, arguing that open models drive global developer adoption of its architecture. Beyond ecosystem lock-in, the deal enables Nvidia to package and sell unused compute capacity to enterprise customers directly through Hugging Face's interface. Huang further highlighted the strategic role of open models in cybersecurity, noting that distributed autonomous defense systems rely on accessible frontier models to protect infrastructure, a capability he deemed vital for both the American and global economies.

The integration of Hugging Face aligns with Nvidia's broader investment strategy in model development, which includes a recent $6 billion deal with coding startup Poolside and over $50 billion infused into AI frontier labs. While the platform promises to stay hardware-agnostic, the combination of Nvidia's dominant silicon market share and control over the primary model repository creates a unified stack that competitors must now navigate. Delangue noted that Nvidia's open models had previously helped defend the platform against cyberattacks when proprietary solutions failed, underscoring the operational synergy driving the merger.

KEY FACTS

ACQUISITION PRICE
$12.93 billion

HOSTED MODELS
3 million

ACTIVE DEVELOPERS
18 million

HOSTED DATASETS
500,000

ANNUALIZED REVENUE
$150 million

TOTAL FUNDING RAISED
$395 million

> WHY IT MATTERS
>
> Practitioners gain assurance that their existing workflows on Hugging Face will remain hardware-agnostic, while gaining potential access to bundled Nvidia compute capacity. However, the consolidation places the primary venue for open-weight model distribution under the control of the dominant AI hardware vendor.

Read the original at TechCrunch →

---

09 MEDIUM IMPACT TechCrunch

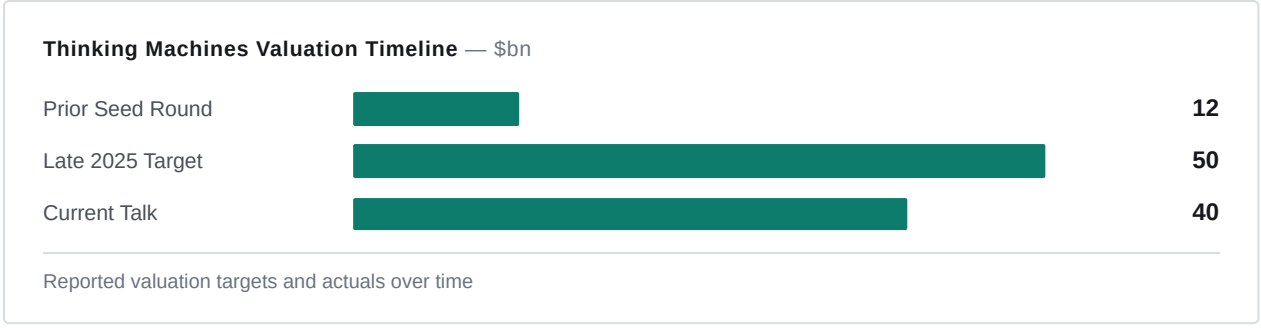
Accel in Talks to Lead $1B for Thinking Machines at a $40B Valuation

> Accel is in talks to lead a $1 billion funding round for Thinking Machines at a $40 billion valuation, marking a step down from the company's previous $50 billion target.

Thinking Machines, the AI laboratory founded in early 2025 by former OpenAI CTO Mira Murati, is negotiating a new capital injection that values the firm at approximately $40 billion. According to reports from The Information and internal sources, existing investor Accel is positioned to lead this $1 billion raise. This proposed valuation represents a contraction from the $50 billion figure the startup reportedly pursued in late 2025, suggesting a market correction or a strategic pivot in pricing expectations despite the company's rapid growth trajectory.

Financial disclosures indicate an annual revenue run rate exceeding $100 million, driven primarily by the July launch of Inkling. This open-weight model monetizes through usage-based compute fees charged for adapting models on proprietary data via the company's Tinker platform. At the reported $40 billion valuation, the company trades at a revenue multiple significantly higher than typical software benchmarks, reflecting investor confidence in the underlying technology rather than current cash flow. The previous financing event, a $2 billion seed round led by Andreessen Horowitz with participation from Nvidia, GV, Lightspeed, and Conviction Partners, had established a $12 billion valuation based largely on the pedigree of Murati and her team of former OpenAI researchers.

The organizational landscape at Thinking Machines has shifted since its inception. While the initial funding round capitalized on the arrival of high-profile talent, the company has subsequently experienced notable departures. Co-founders Lilian Weng and Luke Metz have returned to OpenAI, altering the original founding composition that attracted early capital. Neither Accel nor Thinking Machines provided immediate comment regarding the ongoing negotiations. If completed, this round would solidify Thinking Machines' status as one of the most heavily capitalized private AI entities, even as it accepts a lower valuation ceiling than previously anticipated.



KEY FACTS

PROPOSED VALUATION
$40 billion

FUNDRAISE SIZE
$1 billion

LEAD INVESTOR
Accel

REVENUE RUN RATE
$100 million

PREVIOUS VALUATION TARGET
$50 billion

PRIOR SEED VALUATION
$12 billion

> WHY IT MATTERS
>
> Practitioners should note the divergence between Thinking Machines' $100 million revenue run rate and its $40 billion valuation, signaling that market pricing for foundational labs remains decoupled from traditional revenue multiples. The return of key co-founders to OpenAI may also influence future model development roadmaps and ecosystem alignment.

Read the original at TechCrunch →

---

10 HIGH IMPACT TechCrunch

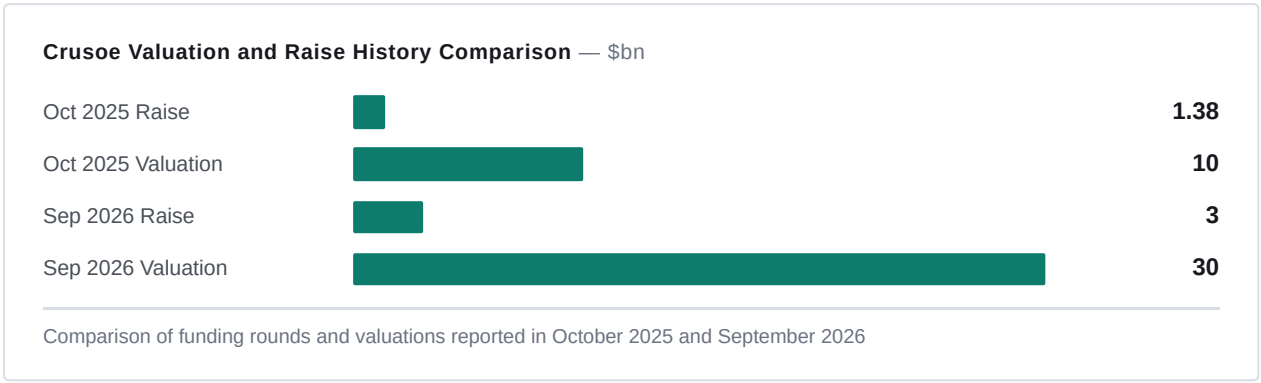
Crusoe Raises $3B at $30B as Data-Center Money Keeps Compounding

> Data center developer Crusoe has secured $3 billion in new funding at a $30 billion valuation, driven by a massive $13 billion infrastructure contract with Jane Street.

Crusoe, originally launched in 2018 as a crypto mining operation utilizing flared natural gas, has pivoted to become a hyperscale AI infrastructure provider for clients including Meta, Microsoft, OpenAI, and Oracle. The company's latest capital injection, reported by Bloomberg, values the firm at $30 billion, a tripling of its $10 billion valuation recorded only ten months prior during a $1.38 billion raise last October. This rapid appreciation underscores the intensifying capital competition for data center capacity capable of supporting large-scale AI workloads.

The primary catalyst for this valuation jump is a newly signed five-year cloud contract with quantitative trading firm Jane Street, valued at $13 billion. Under this agreement, Crusoe will supply GPUs and dedicated AI infrastructure, marking one of the largest single-deal commitments in the sector to date. The fresh $3 billion round was co-led by Atreides Management and Valor Equity Partners, with participation from Mubadala Capital, the asset management arm of Abu Dhabi's sovereign wealth fund. These investors are betting on Crusoe's ability to scale physical infrastructure faster than traditional cloud providers can deploy new regions.

With this liquidity event, Crusoe is now positioning itself for a public market exit. Axios reported last month that the company has engaged Goldman Sachs and Morgan Stanley to discuss a potential near-term IPO. For technical teams relying on third-party compute, this signals a maturing market where specialized infrastructure providers are achieving parity with hyperscalers in terms of financial stability and deployment scale. The shift from a niche flared-gas miner to a $30 billion entity highlights the extreme premium currently placed on immediate GPU availability over long-term energy efficiency narratives alone.



KEY FACTS

LATEST RAISE
$3 billion

POST-MONEY VALUATION
$30 billion

JANE STREET CONTRACT VALUE
$13 billion

PREVIOUS VALUATION
$10 billion

PREVIOUS RAISE AMOUNT
$1.38 billion

TIME SINCE PREVIOUS ROUND
10 months

WHY IT MATTERS

The emergence of well-capitalized, specialized infrastructure providers like Crusoe offers an alternative procurement channel for GPU clusters outside the dominant hyperscalers. Practitioners should evaluate these vendors for large-scale deployments where dedicated capacity and specific energy sourcing models may provide availability advantages.

Read the original at TechCrunch →

---

11 HIGH IMPACT TechCrunch

Meta Will Pay to Watch How You Use Its New AI Model

> Meta is inverting the standard opt-out privacy model by offering explicit financial compensation for users who allow their prompts and outputs to train future versions of its Muse Spark agent.

Meta has introduced a contributor pricing tier for its new Muse Spark model, designed for coding and operational agents, that reduces API costs by approximately 95% in exchange for full data access. Under the standard agreement, 1 million input tokens cost $1.25 and 1 million output tokens cost $4.25. The contributor model slashes these rates to 10 cents and 20 cents respectively. This pricing structure explicitly monetizes the data rights typically retained by enterprise clients, targeting organizations willing to trade data retention governance for significantly lower operational expenses during prototyping and integration testing.

The move addresses a critical bottleneck in developing agentic tools: the scarcity of high-quality digital traces from professional workflows. While consumer-facing coding agents like Claude Code improved rapidly between April 2025 and October 2025 by defaulting to session storage for reinforcement learning, enterprise adoption remains stalled by data privacy concerns. Princeton professor Arvind Narayanan notes that large companies currently accept 10x to 20x higher costs on token-billed Enterprise plans specifically to prevent data usage for model training. Meta's framework attempts to break this deadlock by making the cost of data privacy transparent and optional rather than mandatory.

This strategy emerges amidst intensifying price competition among frontier labs. Anthropic recently lowered costs for processing cached tokens with its Fable and Mythos models, while OpenAI implemented major price cuts on its latest models in late July. Meta's approach differentiates itself by tying price reductions directly to data contribution rather than general market adjustments. The company's pricing guide positions the contributor tier as a mechanism to lower barriers for scaling experiments where training on user data is acceptable, potentially forcing enterprises to rigorously classify which portions of their workflow data are truly proprietary versus shareable.

Meta's internal struggles to secure training data provide context for this external incentive program. An earlier initiative to track employee computer usage faced wide internal criticism and was paused in June 2026. By shifting the data collection burden to external volunteers compensated via API discounts, Meta bypasses internal friction while acquiring the complex interaction logs necessary to refine agentic behaviors. The model does not respond to questions about the specific mechanics of data anonymization or retention periods within the contributor agreement.

KEY FACTS

STANDARD INPUT COST
$1.25 per 1M tokens

CONTRIBUTOR INPUT COST
$0.10 per 1M tokens

STANDARD OUTPUT COST
$4.25 per 1M tokens

CONTRIBUTOR OUTPUT COST
$0.20 per 1M tokens

DISCOUNT RATE
~95%

MODEL NAME
Muse Spark

> **WHY IT MATTERS**
>
> Practitioners building agentic systems must now evaluate whether the 95% cost reduction justifies the risk of exposing proprietary workflows to model training, fundamentally altering the ROI calculation for pilot projects.

Read the original at TechCrunch →

Sources

01 OpenAI Ships GPT-6 Astra: 99.9% on ARC-AGI-3 and a Head-On Fight with Claude Fable
https://simonwillison.net/2026/Sep/3/gpt6-astra/

02 Meta's Muse Spark 1.3 Ranks Third on the Intelligence Index — Fourth Release in Five Months
https://the-decoder.com/meta-closes-in-on-the-top-with-muse-spark-1-3-and-undercuts-rivals-on-price

03 Compile by Training Turns Plain-Language Specs into Local Neural Functions
https://arxiv.org/abs/2609.04199

04 When the LLM Judge Is the Instrument: Preregistered Audit Finds Unstable Measurements
https://arxiv.org/abs/2609.04198

05 Hugging Face's funes Gives Coding Agents a Durable Memory You Own
https://huggingface.co/blog/funes

06 Abliteration.ai Turns Guardrail Removal into a Commercial API Business
https://techcrunch.com/2026/09/03/abliteration-ai-is-making-a-business-out-of-removing-ai-guardrails/

07 Google's WeatherNext 3 Brings Hourly Satellite-Driven Forecasts to Search and Maps
https://deepmind.google/blog/introducing-weathernext-3-our-most-advanced-and-accurate-global-weather-ai-model/

08 Nvidia Confirms the Deal Is Done: Hugging Face Sells for $12.9 Billion
https://techcrunch.com/2026/09/03/nvidia-confirms-it-will-buy-hugging-face-for-12-9-billion/

09 Accel in Talks to Lead $1B for Thinking Machines at a $40B Valuation
https://techcrunch.com/2026/09/03/accel-reportedly-in-talks-to-lead-1b-round-for-thinking-machines-at-40b-valuation/

10 Crusoe Raises $3B at $30B as Data-Center Money Keeps Compounding
https://techcrunch.com/2026/09/03/crusoe-reportedly-raises-3b-at-a-30b-valuation/

11 Meta Will Pay to Watch How You Use Its New AI Model
https://techcrunch.com/2026/09/03/meta-is-paying-to-peek-at-how-you-use-their-latest-ai-model/

About this document. Every story in the 4 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 2026-09-04T05:35:24Z · model qwen3.5:397b-cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.