

The Long Read

1 October 2026

11

STORIES

3

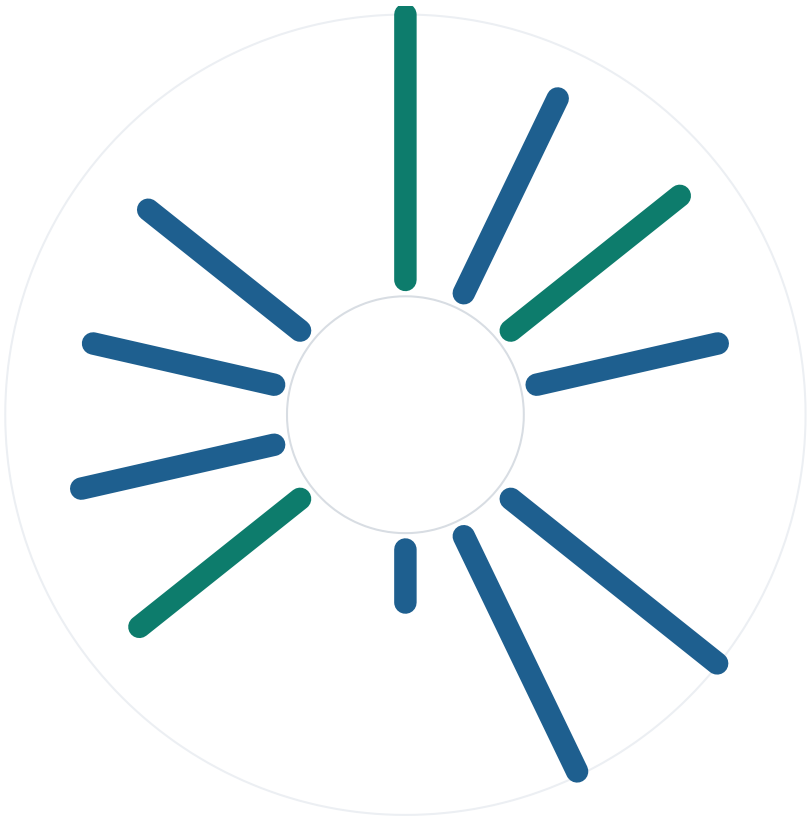
SECTIONS

2881

WORDS

3

HIGH IMPACT



● 3 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

11 stories · 2881 words · 3 sections · 3 charted



AI Models & Research 4 stories · 1142 words



AI Tools & Ecosystem 3 stories · 738 words



AI Applications & Industry 4 stories · 1001 words



AI Models & Research

- 01 ● Google Launches Gemini 4 Argon, Its New Frontier Model for Coding, Enterprise Work and Cyber Defense
- 02 ● SynthID Bio: Watermarks That Survive Into the Physical Protein
- 03 ● Nearly a Third of New Web Text Is Now AI-Generated, and the Pipeline Keeps Drinking It
- 04 ● Semifactual Credit: RLVR Models Still Bend to Prompt Details That Shouldn't Matter

AI Tools & Ecosystem

- 05 ● OpenAI Ships the Decisions API, the Frontier Lab Adopts the System-One Playbook

- 06 ● Hugging Face Debuts an Open Leaderboard for Multilingual TTS and Voice Cloning
- 07 ● Photo Scrubber: Blur Faces and Strip Metadata Locally Before You Post

AI Applications & Industry

- 08 ● Reddit Kills RSS and Its Public API, Blaming the Same AI Bots It Now Charges
- 09 ● DoorDash Launches an Agent You Can Text to Order Dinner
- 10 ● Cerebras' Andrew Feldman Takes the Can-AI-Keep-Scaling Question to Disrupt
- 11 ● Restate Raises 20 Million as Agent Workloads Make Durable Infrastructure a Need, Not a Nice-To-Have

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.

AI Models & Research

4 stories ● 2 high ● 2 medium

01 **HIGH IMPACT** Google

Google Launches Gemini 4 Argon, Its New Frontier Model for Coding, Enterprise Work and Cyber Defense

Google has begun a phased release of Gemini 4 Argon, a frontier model built for long-horizon coding, enterprise knowledge work, and cyber defense, initially limited to trusted defenders under the Fairwind Program.

Gemini 4 Argon is rolling out first to a set of trusted cyber defenders through Google's Fairwind Program, with broader availability to paid API customers and Google AI Ultra subscribers to follow. Pricing is set at \$2 per million input tokens and \$10 per million output tokens, with cached input tokens at 95% off the input price. The model expands output limits from 64K to 1M tokens, which Google says gives it headroom to generate hundreds of thousands of tokens in a single trajectory for deeper reasoning on complex problems.

On benchmarks, Argon sets a new state of the art on DeepSWE v1.1 at 77.9% for long-horizon software engineering tasks, ranks first on AutomationBench at 51.3%, and leads on LVBench for long video understanding at 91.7%. Google also reports leading results on the Vals Index, Vals Finance Agent v2, and Harvey's Legal Agent Benchmark. In cybersecurity, Argon ties for first on CWE-bench v1 at 68%, building on Gemini 3.8 Flash Cyber's performance on CWE-bench v0. Wiz is using Argon through its Scan for Good initiative, where the model uncovered a critical vulnerability in healthcare software that earlier frontier models missed.

Internal Google deployments show concrete engineering impact. Argon agents are migrating C/C++ codebases to Rust, including up to 800K+ lines for the Fuchsia Zircon kernel, with rigorous auditing before production rollout. For libgav1, agents replaced 32K lines of SIMD code with safe Rust that the compiler vectorizes automatically, producing a memory-safe video decoder that runs 2.7x faster than the existing Rust port with identical output. In quantum computing, Argon beat a published baseline for spacetime resource optimization by 40% in minutes. A fleet-wide memory optimization effort freed over 300 TiB of memory, with estimated total savings of 500 TiB to 1 PiB.

Safety work is proceeding in parallel. Google is engaged in the U.S. government's voluntary pre-release model access process, and is strengthening safeguards against misuse, indirect prompt injection, and misalignment. Argon leads on Gray Swan's Indirect Prompt Injection benchmark. For trusted defenders and internal teams, Google is releasing Argon without cyber guardrails to enable full defensive capabilities.

KEY FACTS

INPUT PRICE

\$2 per million tokens

OUTPUT TOKEN LIMIT

1M tokens

CWE-BENCH V1

68%

OUTPUT PRICE

\$10 per million tokens

DEEPSWE V1.1

77.9%

LVBENCH

91.7%

WHY IT MATTERS

The 1M-token output limit and long-horizon reasoning change what is feasible in agentic workflows, while the cyber-defense focus and phased release signal that frontier capability is increasingly gated by safety and trust rather than raw model quality.

[Read the original at Google](#) →

02 **MEDIUM IMPACT** Google DeepMind

SynthID Bio: Watermarks That Survive Into the Physical Protein

Google DeepMind has demonstrated a proof-of-concept watermark that survives translation from digital protein design into the physical synthesized molecule without degrading biological function.

SynthID Bio is a family of watermarking methods for synthetic biology that embeds an imperceptible signature directly into biological code. For protein sequences, it subtly guides amino acid choice; for predicted 3D structures, it adjusts atomic coordinates. The result is a detectable signal that persists in the physical protein, not just the digital artifact. In wet-lab testing across three target proteins — VEGF-A, the SARS-CoV-2 spike protein RBD, and PD-L1 — watermarked binder designs produced with AlphaProteo and a SynthID Bio-enabled ProteinMPNN matched the hit rate, binding affinity, and natural sequence diversity of unwatermarked versions. DeepMind describes these as the first watermarked and biologically functional protein binders.

For structure prediction, the team fine-tuned a small part of AlphaFold 3's diffusion network so the watermark lives in the model weights themselves. Predicted 3D coordinates then carry a detectable signature regardless of who runs the model. DeepMind reports near-perfect detectability while preserving AlphaFold 3 prediction accuracy and key structural feature distributions, with robustness against digital noise and minor coordinate changes.

The biosecurity motivation is concrete. DNA synthesis providers screen orders against databases of known threats, but AI-generated sequences can resemble nothing in those databases, forcing exhaustive manual review. A verifiable watermark lets a screener confirm an order originated from a trusted model with built-in safeguards. The same signal could flag AI-generated entries in public databases such as the Protein Data Bank, UniProt, and GenBank. DeepMind is publishing the methods paper, open-sourcing the code and in vitro data, and releasing weights to the research community. Ongoing work with the Hie lab at Stanford and Arc Institute extends the approach to Evo 2-designed bacteriophage genomes, with early lab tests confirming functional watermarked phages.

KEY FACTS

TARGET PROTEINS IN WET-LAB TESTING

VEGF-A, SARS-CoV-2 spike RBD, PD-L1

STRUCTURE MODEL

AlphaFold 3 diffusion network fine-tune

PARTNER

Hie lab at Stanford and Arc Institute (Evo 2 bacteriophage)

WATERMARKING METHODS

Sequence amino-acid guidance and 3D coordinate adjustment

RELEASE

Methods paper, open-source code, in vitro data, weights

WHY IT MATTERS

For anyone shipping AI-designed proteins or structures, provenance is becoming a screening and database-integrity requirement, and SynthID Bio offers an open, model-level mechanism to provide it without a separate post-processing step.

[Read the original at Google DeepMind →](#)

03

HIGH IMPACT

arXiv.org

Nearly a Third of New Web Text Is Now AI-Generated, and the Pipeline Keeps Drinking It

Wild AI-generated web text now makes up nearly a third of filtered pretraining tokens, and a new 800-model scaling study shows its value to human-text loss flips from benefit to harm as more is added.

After FineWeb quality filtering, 27.5% of tokens from June 2026 web data are labeled AI-generated by Pangram, rising to 31.1% by August. Unlike synthetic data or model-collapse setups, this wild AI text comes from many models, is written for human readers, and arrives unlabeled in pretraining corpora. The authors release WildAI, an 83B-token corpus with AI, topic, and format labels, along with all 800 models and code.

To measure the effect, the team pretrained 800 language models while varying the ratio of added AI tokens to human tokens, then fit scaling laws to held-out losses on both human and AI-generated text. For data-starved models, adding AI tokens initially lowers loss on human text, but the benefit saturates and quickly reverses into harm as more are added. For models already trained on high budgets of human text, AI tokens raise loss almost immediately, while the same number of fresh human tokens keeps lowering it. Existing scaling laws such as Hoffman et al. (2022) fail to predict this behavior.

The authors propose a new scaling law with separate benefit and harm terms that allows the value of an AI token to change sign while reducing to Chinchilla in the absence of AI text. Fit on smaller models, the law predicts the effect of AI text on held-out human-text loss for models up to 3.6x larger with 41% lower error than the best existing law over all AI ratios.

Their recommendations: filter AI text when the target is human text, repeat human text before expanding the training dataset with AI-generated web text, and report validation loss on human and AI text separately. AI text remains valuable when the target is AI text.

AI-generated share of filtered web tokens — %

June 2026		27.5
August 2026		31.1

Share of FineWeb-filtered web tokens labeled AI-generated by Pangram · +13%

KEY FACTS

AI-GENERATED TOKENS, JUNE 2026

27.5%

MODELS PRETRAINED

800

PREDICTION ERROR REDUCTION VS BEST EXISTING
LAW

41%

AI-GENERATED TOKENS, AUGUST 2026

31.1%

WILDAI CORPUS SIZE

83B tokens

EXTRAPOLATION RANGE OF NEW SCALING LAW

3.6x larger models

WHY IT MATTERS

Teams scraping the open web for pretraining data are now almost certainly ingesting substantial AI-generated text, and the study gives a concrete, quantified reason to filter it or repeat human data instead — while warning that standard scaling laws will mislead you about the damage.

[Read the original at arXiv.org](#) →

04

MEDIUM IMPACT

[arXiv.org](#)

Semifactual Credit: RLVR Models Still Bend to Prompt Details That Shouldn't Matter

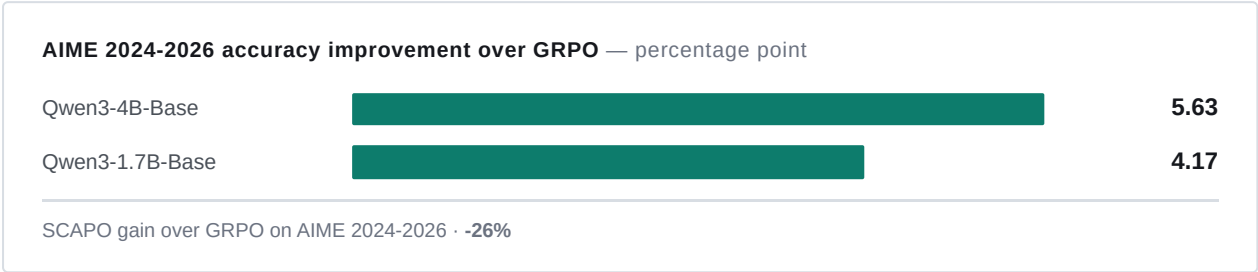
RLVR models still shift their predictions when task-irrelevant prompt details change, and a new causally inspired GRPO variant shows that penalizing those unstable tokens during training improves both accuracy and out-of-distribution generalization.

The paper identifies a concrete failure mode in RLVR-trained LLMs: predictions remain sensitive to semifactual prompt interventions that preserve the underlying problem and its answer. Token-level sensitivity varies substantially across responses, and the authors show that suppressing high-drift token candidates at decoding time improves reasoning accuracy without any weight updates. This points to a structural problem in Group Relative Policy Optimization (GRPO), which assigns the same outcome-derived advantage to every response token, potentially reinforcing spurious dependence on irrelevant prompt features alongside useful reasoning.

The proposed method, Semifactual Credit-Augmented Policy Optimization (SCAPO), modifies GRPO's token-level credit assignment. It measures token probability drift for fixed responses under semifactual interventions, normalizes these stability scores, and reduces advantages for relatively unstable tokens during early training. Stability alone earns no additional credit — the mechanism is purely subtractive for unstable tokens. The approach is evaluated on Qwen3-4B-Base and Qwen3-1.7B-Base.

On AIME 2024-2026, SCAPO improves accuracy over GRPO by 5.63 percentage points on Qwen3-4B-Base and 4.17 percentage points on Qwen3-1.7B-Base. At both model scales, SCAPO achieves the best results on most evaluated mathematics benchmarks and on all evaluated out-of-distribution benchmarks among the compared methods. Code is available via the paper's arXiv page.

The work is a training-time intervention, not a new architecture or scale-up. Its novelty lies in using semifactual stability as a per-token training signal rather than as a decoding heuristic, and in demonstrating that this signal transfers to out-of-distribution settings.



KEY FACTS

AIME 2024-2026 GAIN OVER GRPO (QWEN3-4B-BASE)

5.63 percentage points

BASE MODELS EVALUATED

Qwen3-4B-Base, Qwen3-1.7B-Base

CODE

Available via arXiv page

AIME 2024-2026 GAIN OVER GRPO (QWEN3-1.7B-BASE)

4.17 percentage points

METHOD

Semifactual Credit-Augmented Policy Optimization (SCAPO)

WHY IT MATTERS

If you are running RLVR pipelines with GRPO, this identifies a likely source of spurious prompt sensitivity in your checkpoints and offers a drop-in credit-assignment change that improves both in-distribution math accuracy and out-of-distribution generalization on Qwen3-scale models.

[Read the original at arXiv.org](#) →

AI Tools & Ecosystem

3 stories ● 3 medium

05 MEDIUM IMPACT TechCrunch

OpenAI Ships the Decisions API, the Frontier Lab Adopts the System-One Playbook

OpenAI has entered the fast-decision-model market with a limited-preview Decisions API that mirrors TypeSafe AI's Jev, signaling that cheap, constrained classifiers—not general LLMs—are becoming the default control layer for agentic software.

At Dev Day on Tuesday, Sam Altman announced the Decisions API in an aside: developers give OpenAI's Luna model a predefined set of options—image categories, agent behaviors—and the model outputs a choice. Altman framed the tradeoff explicitly: "By focusing the model on that choice, we can make it extremely fast while keeping capabilities like image understanding, broad language support, and safety protections." The product is a limited preview, and TechCrunch has not yet seen developers benchmark it.

The obvious reference point is Jev, released earlier this month by TypeSafe AI. Jev is a classifier built on an LLM that emits probabilities over a fixed choice set at low cost and high speed. TypeSafe CEO Diogo Almeida, a former OpenAI engineer who co-invented reinforcement learning, responded to the announcement by joking about "the beginning of the clone wars" and suggested OpenAI's interest is "a sign...that building in a System One compatible way is the future." TypeSafe uses "System One" for fast, intuitive inference and "System 2" for deliberate reasoning. Almeida's stated moat is synthetic data that produces statistically calibrated outputs: "Fast and cheap is very easy... If you want it really fast and cheap, use dice, right? Intelligence is the hard part."

The concrete near-term application is agent oversight. After a series of incidents where its agents misbehaved on the open internet, OpenAI added a security measure that uses a separate model to watch for bad actions at "significant compute cost." Shapor Naghibzadeh, who leads QueryStory, built a hackathon demo last weekend that runs Jev on every agentic action, blocking high-confidence bad actions, flagging others, and permitting the rest. His cost comparison: monitoring of that kind runs \$2.94 with Jev versus \$372 with a frontier LLM. That gap is what makes per-action review economically viable.

What is new here is not the technique—other startups are shipping similar models—but the endorsement. A frontier lab adopting the System One playbook validates the argument that general LLMs are the wrong tool for many software control paths. The open question is calibration: how well each decision model's probabilities track real-world outcomes.

Agent monitoring cost: Jev vs frontier LLM — \$

Jev		2.94
Frontier LLM		372

Cost to monitor agentic actions in a hackathon demo · 126.5× higher

KEY FACTS

JEV MONITORING COST

\$2.94

DECISIONS API AVAILABILITY

Limited preview

TYPESAFE CEO

Diogo Almeida, former OpenAI engineer

FRONTIER LLM MONITORING COST

\$372

JEV RELEASE TIMING

Earlier this month (September 2026)

WHY IT MATTERS

If decision models are cheap enough to run on every agentic action, builders can add a review layer that catches misbehavior without the compute bill of a frontier LLM—changing the economics of agent reliability and safety monitoring.

[Read the original at TechCrunch →](#)

06

MEDIUM IMPACT

huggingface.co

Hugging Face Debuts an Open Leaderboard for Multilingual TTS and Voice Cloning

Hugging Face has shipped an objective-metrics leaderboard for open TTS models, cutting evaluation time from weeks of human voting to a couple of hours.

The Open TTS Leaderboard evaluates open-source text-to-speech models on four complementary axes: intelligibility via WER/CER against Qwen3 ASR transcripts, speed via inverse real-time factor (RTFx) for batched H200 inference and time-to-first-audio (TTFA) for streaming batch-size-1 latency, and speaker similarity via cosine similarity between WavLM embeddings of generated audio and reference clips. The default view ranks models by macro-average WER on the English splits of Seed TTS Eval and CV3 Eval (zero shot), with hexgrad/Kokoro-82M, Supertone/supertonic-3, and fishaudio/s2-pro leading that ranking.

The leaderboard is explicitly multilingual: users can toggle languages to rank on CV3 Eval, with CER reported for Chinese, Japanese, and Korean. k2-fsa/OmniVoice, fishaudio/s2-pro, and FunAudioLLM/Fun-CosyVoice3-0.5B-2512 are flagged as strong multilingual performers. A voice-cloning toggle adds a SIM column and Pareto plots for the SIM/inference/size tradeoff; models such as bosonai/higgs-tts-3-4b and openbmb/VoxCPM2 show improved average WER when a reference audio is provided. A separate Streaming tab ranks models by median TTFA on 50 CV3-Eval English prompts after three warm-up runs, with kyutai/pocket-tts highlighted for GPU and CPU streaming.

The motivation is the mismatch between release velocity and evaluation capacity. As of Sep 30, 2026, the Hub hosts more than 8K TTS models, yet only 16 of 92 models on Artificial Analysis are open-weights, with a

similar skew on Voice Arena. Arena-style leaderboards require hosting open models and collecting votes over weeks, and they cannot guarantee consistent voter criteria over time. Objective metrics drop evaluation to roughly two hours per model. The authors are careful to state this does not replace human preference: WER proxies intelligibility and SIM proxies voice identity, but neither measures naturalness or expressiveness. A Listen tab lets users compare generated outputs and vote, with plans to feed community votes back into the leaderboard. Evaluation scripts will be open-sourced, following the Open ASR Leaderboard repo pattern.

What is genuinely new is the combination of multilingual coverage, voice-cloning evaluation, streaming latency measurement, and open-model focus in a single public leaderboard with reproducible objective metrics. Prior arena-style efforts covered preference ranking but underrepresented open weights and could not scale with release cadence.

KEY FACTS

TTS MODELS ON HUB (SEP 30, 2026)

8K+

EVALUATION TIME, OBJECTIVE METRICS

~2 hours

STREAMING BENCHMARK PROMPTS

50 English prompts from CV3-Eval

OPEN-WEIGHTS MODELS ON ARTIFICIAL ANALYSIS

16 of 92

EVALUATION TIME, ARENA VOTING

~2 weeks

ASR USED FOR WER/CER

Qwen3 ASR

WHY IT MATTERS

Practitioners evaluating open TTS models for multilingual or voice-cloning deployments get a reproducible, fast first-pass filter before committing to costly human preference testing, with streaming TTFA data directly relevant to voice-agent latency budgets.

[Read the original at huggingface.co](#) →

07

MEDIUM IMPACT

Simon Willison's Weblog

Photo Scrubber: Blur Faces and Strip Metadata Locally Before You Post

Simon Willison's weblog covers Photo Scrubber, a tool for blurring faces and stripping metadata locally before posting.

Source not retrievable. This entry is written from the headline and the editor's summary only — the publisher blocked automated retrieval (extracted only 97 words (paywall/consent wall?)). Follow the link for the full report.

The post highlights a local-first tool that blurs faces and removes EXIF metadata in one pass before a photo leaves the user's machine. The full article could not be retrieved, so no further details are available. The headline and summary indicate the tool operates entirely on-device and targets pre-publication privacy.

WHY IT MATTERS

If the tool works as described, it would give users a simple way to reduce two common privacy leaks before sharing photos.

[Read the original at Simon Willison's Weblog](#) →

AI Applications & Industry

4 stories ● 1 high ● 3 medium

08 **HIGH IMPACT** TechCrunch

Reddit Kills RSS and Its Public API, Blaming the Same AI Bots It Now Charges

Reddit is shutting down RSS feeds on November 13 and ending public API access by March 2027, pushing all programmatic data access behind commercial deals.

Reddit announced Wednesday that RSS support will cease on November 13, describing the format as a "common surface for large-scale scraping and automated abuse." The company acknowledged RSS as "a beloved part of the open web" but offered no replacement for feeds used outside moderators' own communities. For moderators relying on RSS for alerts, Reddit recommends migrating to the Discord Relay Devvit app before the shutdown date.

The larger change is the end of public API access by March 2027. This affects any tool using programmatic access to Reddit conversations: social listening products, researcher tooling, and AI assistants that currently pull Reddit content to answer questions. After the cutoff, those systems will need commercial data deals with Reddit. Developers building approved third-party apps and bots must register with the company before January 12, 2027, or lose API access.

The timing is not incidental. Reddit's "other revenue" beyond advertising grew 24% year-over-year to \$43 million in Q2, driven by AI licensing deals. The company is now closing the free paths to the same data it sells. Reddit is also adding safeguards to Old Reddit, limiting access to logged-in moderators and users who have used the classic interface within the last 6 months — a window Reddit extended from 90 days as the news went out — again citing "abusive scraping and automated traffic."

Moderators have already warned that losing RSS will make communities harder to manage, while users said RSS was how they preferred to consume Reddit content. The company's position is clear: free, standardized access is being retired in favor of paid, contracted access.

KEY FACTS

RSS SHUTDOWN DATE	PUBLIC API SHUTDOWN DATE
November 13	March 2027
DEVELOPER REGISTRATION DEADLINE	Q2 OTHER REVENUE
January 12, 2027	\$43 million
OTHER REVENUE YOY GROWTH	OLD REDDIT RECENT-ACCESS WINDOW
24%	6 months

Any AI assistant, research pipeline, or social listening tool that currently pulls Reddit data through RSS or the public API has a hard deadline to either negotiate a commercial deal or lose access. Teams should audit Reddit dependencies now, since registration for approved third-party apps closes January 12, 2027.

[Read the original at TechCrunch →](#)

09

MEDIUM IMPACT

TechCrunch

DoorDash Launches an Agent You Can Text to Order Dinner

DoorDash is moving food ordering out of its app and into Apple Messages with a text-to-order AI agent that interprets prompts like "order my usual."

DoorDash announced on Wednesday that it is launching a text-to-order AI agent accessible through Apple Messages. Users can send prompts such as "order my usual," and the agent resolves that to the user's Friday night order. The system also accepts requests for a specific dish or a local recommendation, then searches nearby restaurants and suggests a cart based on the prompt. When it recommends food, the agent can text photos of the dishes. For group orders, DoorDash says the agent can handle mixed dietary preferences and different quantities within a single order.

The launch positions DoorDash against Uber Eats and Grubhub in the race to add agentic ordering surfaces. It also lands amid a broader industry push toward personal AI agents that complete tasks without requiring users to navigate apps or websites. DoorDash is opening a U.S. waitlist for the feature rather than rolling it out broadly, which suggests a controlled test phase.

Separately, DoorDash said it will begin testing delivery drones with select restaurants in Northern California. The announcement did not include a timeline, restaurant names, or technical specifications for the drone program.

For practitioners, the notable detail is the interaction model: the agent operates inside an existing messaging surface rather than a DoorDash-owned interface. That shifts the integration problem from building a conversational UI to handling intent resolution, cart construction, and order confirmation across a third-party channel. The source does not disclose the underlying model, latency, or how the agent handles payment and address confirmation, so the technical substance remains thin.

KEY FACTS

LAUNCH CHANNEL

Apple Messages

AVAILABILITY

U.S. waitlist

DRONE TEST REGION

Northern California

ANNOUNCED

Wednesday

WHY IT MATTERS

It signals that consumer ordering flows are moving into messaging surfaces where intent parsing, cart assembly, and multi-user constraint handling become the core engineering problem rather than app navigation.

[Read the original at TechCrunch →](#)

Cerebras' Andrew Feldman Takes the Can-AI-Keep-Scaling Question to Disrupt

Cerebras Systems is bringing wafer-scale AI compute to TechCrunch Disrupt 2026 with a session on whether AI can keep scaling, backed by a \$5.5 billion IPO and a 750-megawatt OpenAI deployment agreement.

Andrew Feldman, CEO and co-founder of Cerebras Systems, will take the Disrupt Stage at TechCrunch Disrupt 2026 for a session titled "Can AI Keep Scaling?" The talk, scheduled for October 13-15 at Moscone West in San Francisco, will address growing demand for compute, energy, and infrastructure, how Cerebras approaches those constraints, and what happens if conventional hardware reaches its limits.

Cerebras has spent a decade pursuing wafer-scale computing — building a processor on the silicon wafer itself rather than cutting it into individual chips — an architecture designed specifically for AI workloads. The company's recent milestones put that bet in context: a \$5.5 billion IPO in May, a multiyear agreement with OpenAI to deploy 750 megawatts of Cerebras systems from 2026 through 2028, and the August introduction of CS-4, the latest generation of its wafer-scale AI infrastructure.

Infrastructure, not just silicon, is the constraint Feldman will address. Cerebras reported in August that more than 600 megawatts of data center capacity is live or under contract for delivery by the end of 2027, with manufacturing capacity increasing more than tenfold during 2026. The company also plans to bring its first European data center capacity online this year and expand to 200 megawatts there by the end of 2027.

The session is one of 200+ across six industry stages, with more than 10,000 founders, investors, operators, and tech leaders expected, along with 250+ speakers and 300+ exhibiting startups.

KEY FACTS

IPO

\$5.5 billion (May)

DATA CENTER CAPACITY

600+ MW live or under contract by end of 2027

EUROPEAN CAPACITY TARGET

200 MW by end of 2027

OPENAI AGREEMENT

750 MW, 2026-2028

MANUFACTURING SCALE-UP

10x during 2026

LATEST HARDWARE

CS-4, introduced August

WHY IT MATTERS

For teams planning AI infrastructure, Cerebras' wafer-scale approach and its 600MW-plus data center pipeline represent a concrete alternative to conventional GPU-centric scaling — one that is now backed by OpenAI-level deployment commitments rather than just research claims.

[Read the original at TechCrunch →](#)

Restate Raises 20 Million as Agent Workloads Make Durable Infrastructure a Need, Not a Nice-To-Have

Restate has closed a \$20 million Series A to scale its durable workflow engine against Temporal, betting that crash-resilient execution becomes default infrastructure for AI agent workloads.

The Berlin-based startup, co-founded in 2022 by Apache Flink co-creator Stephan Ewen, builds an execution engine that makes multistep workflows resilient to crashes and network interruptions. Ewen said the system was not originally designed for agents, but its guarantees — tracking exactly what a workflow does to keep outcomes reproducible and consistent — map directly onto agentic processes, which run longer and take unpredictable paths compared to traditional software.

Architecturally, Restate differs from competitors by not building on an external database. The company developed its own storage, replication, and redundancy layers, which Ewen said makes the system fast and lightweight relative to heavyweight durability engines. The startup reports closing multiple six- and seven-figure customer contracts in recent months, and counts vibe-coding platform Replit among its users, alongside Fortune 500 customers including in financial services.

The Series A was led by Singular, with participation from Redpoint Ventures and Capital One Ventures. The competitive context is Temporal, founded in 2019, which announced a \$550 million Series E at a \$12.55 billion valuation earlier this month. Restate's founding team includes former Data Artisans and Ververica colleagues Igal Shilman and Till Rohrmann, plus Ahmed Farghal, previously an engineer at Stripe and Meta, who joined as fourth co-founder at the Series A.

The capital will fund a go-to-market team, additional engineers, and a Bay Area office expansion. Ewen's stated thesis is that durable execution becomes a default tool — "like a database" — for companies of any size within a couple of years.

KEY FACTS	
SERIES A	LEAD INVESTOR
\$20 million	Singular
PARTICIPANTS	FOUNDED
Redpoint Ventures, Capital One Ventures	2022
NOTABLE CUSTOMER	TEMPORAL VALUATION (COMPARISON)
Replit	\$12.55 billion

WHY IT MATTERS

As agent workflows become longer-running and less predictable, teams need execution layers that can recover mid-process without manual intervention. Restate's self-contained storage and replication design is a concrete alternative to database-backed durability engines for builders weighing latency, cost, and operational overhead.

[Read the original at TechCrunch](#) →

Sources

- 01 Google Launches Gemini 4 Argon, Its New Frontier Model for Coding, Enterprise Work and Cyber Defense
<https://deepmind.google/blog/gemini-4-argon-our-next-era-of-frontier-intelligence/>
- 02 SynthID Bio: Watermarks That Survive Into the Physical Protein
<https://deepmind.google/blog/introducing-synthid-bio/>
- 03 Nearly a Third of New Web Text Is Now AI-Generated, and the Pipeline Keeps Drinking It
<https://arxiv.org/abs/2609.40295>
- 04 Semifactual Credit: RLVR Models Still Bend to Prompt Details That Shouldn't Matter
<https://arxiv.org/abs/2609.40360>
- 05 OpenAI Ships the Decisions API, the Frontier Lab Adopts the System-One Playbook
<https://techcrunch.com/2026/09/30/openais-jev-clone-could-help-the-frontier-lab-stop-its-swarming-agents/>
- 06 Hugging Face Debuts an Open Leaderboard for Multilingual TTS and Voice Cloning
<https://huggingface.co/blog/open-tts-leaderboard>
- 07 Photo Scrubber: Blur Faces and Strip Metadata Locally Before You Post
<https://simonwillison.net/2026/Sep/29/photo-scrubber/>
- 08 Reddit Kills RSS and Its Public API, Blaming the Same AI Bots It Now Charges
<https://techcrunch.com/2026/09/30/reddit-is-killing-rss-feeds-ending-public-api-access-because-of-ai-bots/>
- 09 DoorDash Launches an Agent You Can Text to Order Dinner
<https://techcrunch.com/2026/09/30/doordash-launches-an-ai-agent-you-can-text-to-order-food/>
- 10 Cerebras' Andrew Feldman Takes the Can-AI-Keep-Scaling Question to Disrupt
<https://techcrunch.com/2026/09/30/cerebras-systems-andrew-feldman-on-whether-ai-can-keep-scaling-at-techcrunch-disrupt-2026/>
- 11 Restate Raises 20 Million as Agent Workloads Make Durable Infrastructure a Need, Not a Nice-To-Have
<https://techcrunch.com/2026/09/30/restate-lands-20m-as-the-need-for-durable-infrastructure-increases-with-ai-agents/>

About this document. Every story in the 1 October 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 2026-10-01T05:34:54Z · model deepseek-v4-pro:cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.