

The Long Read

16 October 2026

12

STORIES

3

SECTIONS

3467

WORDS

5

HIGH IMPACT



● 5 high impact ● 7 medium impact spoke length = depth of coverage

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

The issue at a glance

12 stories · 3467 words · 3 sections · 3 charted



AI Models & Research 4 stories · 1042 words



AI Tools & Ecosystem 3 stories · 1005 words



AI Applications & Industry 5 stories · 1420 words



AI Models & Research

- 01 ● Meta Ships Muse Spark Open: an 816B-Parameter Open-Weight Model Built for Macs
- 02 ● Anthropic Opens the Door: METR Gets Employee-Level Access — and the Company Can't Veto Its Reports
- 03 ● Scaling Indic Pretraining: 2.8 Trillion Tokens Through an Open Data Pipeline
- 04 ● LLMs Prove They Can Design State-of-the-Art Quantum Circuits

AI Tools & Ecosystem

- 05 ● Bluesky's Attie Turns Vibe Coding Into a Group Project — Sites and Games on the AT Protocol

- 06 ● FLUX.3 Image Klein Lands in Diffusers: 5.6B Open Weights With Editing and Layout Control
- 07 ● Choosing a Chunking Strategy for Your RAG Stack: a Decision Tree With Failure Signals

AI Applications & Industry

- 08 ● Anthropic Told Investors: \$16.8B Quarterly Revenue, Up Nearly 50%, IPO in Sight
- 09 ● Targeted Ads Delivered 0-Click Mercenary Spyware to Al Jazeera's Devices
- 10 ● AI Agents Ride Into Rivian's Cars — Anthropic and OpenAI Models Are Next
- 11 ● Amazon's Bee Doubles Down: Two All-Day AI Note-Taking Wearables From \$75
- 12 ● DeepMind Sends Antigravity to the Lab Bench — a New Antibiotic Candidate Out the Other Side

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 MEDIUM IMPACT TechCrunch

Meta Ships Muse Spark Open: an 816B-Parameter Open-Weight Model Built for Macs

Meta has released Muse Spark Open, an 816-billion-parameter open-weight model designed specifically to run locally on Apple hardware.

Meta launched Muse Spark Open on Thursday, describing it as the "most capable" open-weight model in the United States. The model uses 816 billion parameters and a 1 million token context window. It is natively multimodal and targeted at complex coding tasks and agentic labor. Meta's evaluation results claim superior performance over other popular open-weight models, though the source does not include specific benchmark figures.

The defining design choice is local execution on Macs. Running the model requires at least 1 terabyte of RAM, per Engadget, and Meta indicates the best target is the Mac Studio equipped with the M5 Ultra chip. That hardware requirement puts it beyond most consumer machines, but for practitioners with access to a high-end Mac Studio, it offers a path to running a very large open-weight model without cloud inference costs.

Meta paired Muse Spark Open with Muse Code, its AI coding terminal, and deployed it across agentic workflows. One cited deployment involved coding fixes for a Mastodon client, including "bug fixes, new features, [and] full build." The company also used the model with DaVinci Resolve, where it was capable of reviewing and cropping footage. These are concrete demonstrations of the model operating as an agent across desktop applications rather than as a chat-only assistant.

The model is available now through multiple API platforms. Meta did not specify which platforms in the source text. Open-weight models, Meta notes, offer cost efficiency, transparency, and flexibility advantages over closed proprietary models — a familiar argument, but one that carries more weight when the model in question is positioned for local, on-device deployment.

KEY FACTS

PARAMETERS

816B

MINIMUM RAM

1TB

MODALITY

Natively multimodal

CONTEXT WINDOW

1M tokens

TARGET HARDWARE

Mac Studio with M5 Ultra

For teams already invested in Apple silicon, Muse Spark Open offers a route to running an 816B-parameter model locally, avoiding per-token API costs for coding and agentic workloads — provided they can meet the 1TB RAM requirement.

[Read the original at TechCrunch →](#)

02

MEDIUM IMPACT

TechCrunch

Anthropic Opens the Door: METR Gets Employee-Level Access — and the Company Can't Veto Its Reports

Anthropic has given METR employee-level access to its systems and surrendered the right to block the nonprofit's published safety assessments.

Anthropic announced Thursday that six researchers from METR, the nonprofit that tests frontier AI models, will receive employee-level access to independently assess the company's models, safety practices, and risk management. The METR team gets Anthropic-issued laptops, office space, and internal tools roughly on par with what Anthropic's own risk assessment teams use. Anthropic employees are encouraged to share information with them as they would with a colleague working on alignment risk.

The arrangement is the top tier of a three-tier evaluator system Anthropic is building. Bottom-tier evaluators, such as biosecurity nonprofit SecureBio, run narrow reviews and require Anthropic's approval to publish. The middle tier gets more information and can publish without approval. METR sits alone at the top with employee-level access. Anthropic compared the setup to banking, where regulators place resident examiners inside the largest banks.

On publication, Anthropic has given up veto power. The company can require redactions only in narrow, pre-defined categories: customer data, attorney-client privileged material, or information about model architectures. METR may publicly characterize the redactions and say whether it believes they were justified, provided it does not reveal the redacted content.

The announcement fulfills a promise CEO Dario Amodei made in a September essay proposing embedded outside evaluators as a first step toward slowing AI development. OpenAI quickly pledged the same, and Anthropic, OpenAI, Google, Meta, Nvidia, and xAI have since signed a White House accord committing to independent review. The move comes amid a run of disclosed incidents, roughly every week or two since late August, where AI agents acted outside their instructions.

KEY FACTS

METR RESEARCHERS GRANTED ACCESS

6

VETO OVER METR PUBLICATION

None

EVALUATOR TIERS

3

WHITE HOUSE ACCORD SIGNATORIES

Anthropic, OpenAI, Google, Meta, Nvidia, xAI

For practitioners, this signals that frontier-lab safety claims will increasingly face third-party scrutiny with real internal visibility. Teams building agentic systems should expect external audits to become a standard part of deployment at scale, not a one-off compliance exercise.

[Read the original at TechCrunch →](#)

03

HIGH IMPACT

huggingface.co

Scaling Indic Pretraining: 2.8 Trillion Tokens Through an Open Data Pipeline

NVIDIA, IIT Madras, and AI4Bharat are releasing a 2.8-trillion-token synthetic Indic pretraining corpus spanning 22 constitutionally recognized Indian languages, along with the open GPU-accelerated pipeline that produced it.

The corpus, Nemotron-pretraining-Indic-v1, contains 2,601.0 billion native-script tokens and 224.6 billion romanized tokens, counted with the Gemma-4 tokenizer. Twelve languages—Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Odia, Punjabi, Tamil, Telugu, and Urdu—receive both native-script and romanized data, while the remaining ten are native-script only. Source material comes from 15 Nemotron-CC-v2 subsets covering web text, question-answer content, mathematics, code, and other specialized material. Every record retains a source identifier, so outputs trace back to their original English document.

The twelve-language workflow uses Gemma-4-31B-it for document-level translation, quality assessment, and targeted repair. Outputs must score at least 75 out of 100 on an MQM-inspired rubric covering meaning, fluency, terminology, style, and formatting. Feedback-based rewriting recovered 438.4 billion tokens, already included in the corpus total. Romanized text is produced by converting accepted native-script translations into Latin script rather than translating directly from English; initial human comparison found this two-step approach produced fewer hallucinated additions and less residual English. The ten-language workflow uses Bodhan AI's Indic Translate model, with adaptations for Santali (overlapping windows) and Manipuri (Bengali script to Meitei Mayek conversion), and no feedback-based rewriting.

Continued pretraining on Nemotron 3.5 Super VL used a 100-billion-token budget, with 10 billion tokens (10%) drawn from a twelve-language subset of this corpus plus natural Indic text. Scores rose on all five reported Indic benchmarks—MILU, Indic MMLU-Pro, Indic ARC-Challenge, Indic ARC-Easy, and Indic GSM8K—with gains of 4.98 to 9.82 percentage points. The released model also outperforms Nemotron 3 Super on five of six task benchmarks and all four IndiVibe dimensions, though IFEval 1.5 instruction following dipped 2.04 points.

The generation recipe ships as an open-source Nemotron Bharat translation plugin for NeMo Data Designer, including prompts, evaluation rubrics, optional rewriting, document windowing, and script-remapping. The corpus is not yet live; the dataset card will carry the repository ID and license details.

KEY FACTS

CORPUS SIZE

2.8 trillion tokens

ROMANIZED TOKENS

224.6 billion

QUALITY THRESHOLD

75/100 rubric score

NATIVE-SCRIPT TOKENS

2,601.0 billion

LANGUAGES COVERED

22

TOKENS RECOVERED VIA REPAIR

438.4 billion

WHY IT MATTERS

Teams building Indic models get a large, provenance-tracked synthetic corpus plus a reproducible recipe they can run on their own source documents, rather than a one-off data dump. The CPT results suggest synthetic-plus-natural Indic mixtures improve reasoning and knowledge benchmarks, but the 75/100 acceptance threshold and automated judging mean downstream validation is still required.

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

04

HIGH IMPACT

arXiv.org

LLMs Prove They Can Design State-of-the-Art Quantum Circuits

An evolutionary LLM loop with an independent evaluator produced quantum circuits that beat state-of-the-art Toffoli and qubit counts on seven of eight benchmark tasks.

The framework is simple in structure: LLMs propose circuit designs, and an independent evaluator verifies and scores them, feeding results back into the search. The authors applied this loop to tasks spanning numerical computation, chemistry, optimization, quantum simulation, and number theory. On seven of the eight tasks, the discovered circuits significantly reduce both Toffoli and qubit counts relative to the state of the art.

The eighth task is the most interesting failure-turned-result. The search did not beat the incumbent on the usual metrics; instead it discovered a new reversible architecture for prime-field modular inversion. The key difference is how the architecture handles reversibility, which makes divsteps — a classical algorithm previously considered impractical for reversible integer arithmetic — viable. The result substantially improves the Toffoli-qubit trade-off for modular inversion.

What is genuinely new here is not that LLMs can write code or even circuits, but that evaluator-guided co-design can surface algorithmic structures a human search did not. The reversibility constraint in quantum computation is a hard design pressure absent from classical work, and the modular inversion result suggests LLM search can exploit that constraint rather than merely work around it. The paper frames this as opening the door to evaluator-guided LLM co-design of new algorithmic structures for quantum computation, not just optimization of known ones.

KEY FACTS

TASKS WHERE CIRCUITS BEAT STATE OF THE ART

7 of 8

SOURCE

arXiv:2610.22367

SUBMISSION DATE

15 Oct 2026

WHY IT MATTERS

For teams optimizing quantum primitives, an evaluator-scored LLM search loop is a practical way to attack Toffoli and qubit budgets directly, and it can surface architectures that sidestep constraints like reversibility rather than paying their full cost.

[Read the original at arXiv.org](#) →

AI Tools & Ecosystem

3 stories ● 1 high ● 2 medium

05 MEDIUM IMPACT TechCrunch

Bluesky's Attie Turns Vibe Coding Into a Group Project — Sites and Games on the AT Protocol

Bluesky is turning Attie from a feed-building experiment into a general-purpose creation tool whose output ships with built-in distribution across the AT Protocol network.

At TechCrunch Disrupt 2026, Bluesky announced that Attie — its experimental AI platform originally built for custom social feeds — now supports creating websites, quizzes, games, polls, and other interactive experiences. The pitch is explicitly social: CEO Toni Schneider framed it as "social vibe coding or vibe coding for your community," contrasting it with the single-user, solo nature of current vibe coding tools. Creators are connected to their ATProto identity, and Attie will suggest projects based on their public Bluesky profiles. Users can bring their own AI model or use open-weight options.

The distribution angle is the substantive differentiator. COO Rose Wang argued that the unsolved problem in AI-generated apps is not creation but usage: "Everybody is building ... But the big question is: Who's using the thing you're building." Attie's output is not standalone — it plugs into Bluesky's existing network of 47 million registered users. That said, the source notes engagement has declined as some users who tested Bluesky returned to X, which tempers the reach claim.

The announcement also signals protocol-level changes. Private data support on ATProto — described by Schneider as the biggest developer request — is now in alpha testing with a rollout planned by the end of the year. Attie is positioned as both a testing ground for AI ideas and part of a broader push to reinvigorate activity by focusing on developers and small communities. Schneider's framing of the target use case is deliberately narrow: "lots and lots of these fun, small social experiences for your friends," not tools for broadcasting to the entire world.

Leadership context: Schneider, formerly interim CEO after Jay Graber stepped down in March to become chief innovation officer, is now taking charge as Bluesky enters what the article calls the AI era.

KEY FACTS

BLUESKY REGISTERED USERS	PRIVATE DATA STATUS
47 million	Alpha testing
PRIVATE DATA ROLLOUT	ANNOUNCEMENT VENUE
Planned by end of year	TechCrunch Disrupt 2026

For builders on ATProto, Attie lowers the barrier to shipping interactive experiences with an existing distribution channel, while the private-data alpha signals that authenticated, non-public data flows are coming to the protocol by year-end.

[Read the original at TechCrunch →](#)

06

HIGH IMPACT

huggingface.co

FLUX.3 Image Klein Lands in Diffusers: 5.6B Open Weights With Editing and Layout Control

Black Forest Labs' FLUX.3 Image Klein is now in Diffusers as a single 5.6B-parameter model covering text-to-image, multi-reference image editing, and bounding-box layout control.

The release ships two checkpoints in the FLUX.3 Image Klein collection: a time-step-distilled model running 16 steps with no CFG, and a base model running 50 steps with classifier-free guidance. The distilled checkpoint generates a 1024×1024 image in about 2 seconds on an H100. Quantized single-file variants are available in rowwise FP8 and NVFP4, loadable via `from_single_file`; FP8 kernels are fast on compute capability 8.9+ (Ada, Hopper, Blackwell), while NVFP4 targets Blackwell with `flashinfer-python` and `flashinfer-cubin` at matching versions. Peak memory after inference runs 15.3 GiB for NVFP4 on an RTX 6000, 16.27 GiB for FP8 on an H100, and 20.72 GiB for bf16 on an H100.

Architecturally, FLUX.3 is a latent-space flow model using the Flux.2 VAE, which packs 2×2 latent patches into 128-channel tokens, so a 1024×1024 image becomes a 64×64 grid of 4,096 tokens. The text encoder is Qwen3-VL, but instead of its final layer, FLUX.3 stacks hidden states from eight of its 36 layers (4 through 32), producing 20,480-dimensional vectors per text token before projection to width 3072. The transformer runs two stages: five per-mode blocks where text, noisy image, and reference streams attend only to themselves, then 28 joint blocks where all tokens attend across streams. Unlike FLUX.1 and FLUX.2, modulation is per-token-type: noisy image tokens use the current timestep, while text and reference tokens are always modulated at timestep 0. RoPE positions span four axes (t, h, w, l), with references sharing the output's (h, w) coordinates but separated along l, which is what makes `src_bbox/tgt_bbox` editing precise.

Layout control takes a JSON list of regions with `id`, `bbox` on a 0–1000 normalized scale, and `desc`, referenced in the prompt as `.` For edits, regions specify `src_bbox` in the reference and `tgt_bbox` in the output. The pipeline ships as Modular Diffusers, so blocks can be inspected, removed, or inserted without loading weights; the text encoder can be popped out, run once, and freed for repeated generation from cached embeddings. DreamBooth LoRA scripts are included for both text-to-image and instruction-based image editing, with flags for gradient checkpointing, cached latents, and 8-bit Adam.

Peak allocated memory after inference — GiB

nvfp4 (RTX6000)		15.3
fp8r (H100)		16.27
bf16 (H100)		20.72

Peak allocated memory after inference by checkpoint format

KEY FACTS

PARAMETERS

5.6B

BASE INFERENCE

50 steps, CFG

PEAK MEMORY, BF16 AFTER INFERENCE (H100)

20.72 GiB

TEXT ENCODER

Qwen3-VL, 8 of 36 layers stacked

PER-MODE BLOCKS

5

DISTILLED INFERENCE

16 steps, no CFG

1024×1024 GENERATION ON H100

1.93 s

PEAK MEMORY, NVFP4 AFTER INFERENCE (RTX 6000)

15.3 GiB

JOINT BLOCKS

28

WHY IT MATTERS

One open checkpoint now covers generation, multi-reference editing, and layout control, with quantized weights that fit on smaller GPUs and modular internals that make custom training and pipeline surgery practical.

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

07

MEDIUM IMPACT

MachineLearningMastery.com

Choosing a Chunking Strategy for Your RAG Stack: a Decision Tree With Failure Signals

Chunking strategy should be chosen by walking a six-question decision tree tied to document structure and retrieval needs, not by defaulting to fixed-size splits.

The article lays out nine chunking strategies — fixed-size, recursive, document-based, semantic, late, adaptive, hierarchical, LLM-based, and agentic — each tied to a specific assumption about the document. Fixed-size chunking assumes boundary placement doesn't matter because overlap can patch cuts. Document-based chunking assumes headers and sections align with semantic structure. Late chunking assumes a chunk's meaning depends on the whole document. Agentic chunking assumes the collection is varied enough that no single rule handles all of it.

The decision tree runs through six questions in order. First, skip chunking entirely if content fits the context window and covers one topic. Second, if the document has reliable structure — Markdown headings, HTML elements, code functions — use document-based chunking. Third, if chunks need context from elsewhere in the document, such as a legal contract defining a term in section one and using it in section eight, use late chunking with a long-context embedding model. Fourth, if information density varies sharply, use adaptive chunking; if the corpus must answer both broad and specific queries, use hierarchical chunking with coarse parent and finer child chunks. Fifth, if retrieval quality outweighs cost, use LLM-based chunking; if budget is moderate and topic shifts are unpredictable, use semantic chunking on sentence-embedding distance. Sixth, for heterogeneous collections use agentic chunking; for uniform corpora, recursive chunking is the practical default, with fixed-size as a fallback for fast prototypes.

The article also catalogs failure signals and fixes. Chunks consistently cut mid-sentence or mid-table indicate a structure-blind method on structured documents — switch to recursive chunking rather than widening overlap.

Semantic chunking producing wildly inconsistent chunk counts across similar documents signals a miscalibrated similarity threshold. Document-based chunks ranging from one line to several pages mean headers don't track topic boundaries; combine document-based splitting with a recursive pass inside oversized sections. Agentic chunking blowing past compute budget usually means per-document analysis is repeated on every run instead of cached.

What is genuinely new here is not any single technique — all nine strategies are known — but the explicit mapping from document conditions to strategy choice, plus the failure-signal table that turns chunking from a preprocessing afterthought into a diagnosable component of the retrieval pipeline.

KEY FACTS

DECISION TREE QUESTIONS

6

CHUNKING STRATEGIES COVERED

9

FAILURE SIGNALS CATALOGED

5

WHY IT MATTERS

Poor chunk boundaries cause retrieval to miss the passage that answers a query or return chunks lacking surrounding context, and reranking or prompt tuning cannot fully recover what splitting destroyed. The decision tree gives practitioners a concrete starting point and a set of symptoms to watch before blaming the embedding model or generator.

[Read the original at MachineLearningMastery.com](#) →

AI Applications & Industry

5 stories ● 2 high ● 3 medium

08 **HIGH IMPACT** Simon Willison's Weblog

Anthropic Told Investors: \$16.8B Quarterly Revenue, Up Nearly 50%, IPO in Sight

Anthropic's Q3 revenue reached \$16.8bn — up from \$11.5bn in Q2 — and the company says it has materially cut revenue from rival labs distilling its models.

Anthropic hosted investors on Wednesday and disclosed quarterly revenue figures that show accelerating growth: roughly \$4.8bn in Q1, \$11.5bn in Q2, and \$16.8bn in Q3, according to people with knowledge of the matter cited by the Financial Times. The Q3 figure represents nearly 50% quarter-over-quarter growth. In Q3, the company generated \$1.70 in revenue for every \$1 spent on computing power, and adjusted operating margins excluding stock-based compensation were 11%.

The most consequential disclosure concerns distillation. Anthropic told investors that distillation — the practice of rival AI labs training their models on the outputs of more advanced models — accounted for 5% to 7% of revenue, and that the company cut it back "materially" in Q3. Four people with direct knowledge of the matter relayed this to the FT. If distillation represented 5% of Q2 revenue before the cutback, that implies unnamed rival labs spent over half a billion dollars in a single quarter distilling Anthropic's models.

Simon Willison notes that Anthropic appears to have stopped leaning on annualized run-rate revenue in these investor communications, a framing the company had used previously. The shift to raw quarterly figures is itself a signal about how the company now wants its trajectory read ahead of a possible IPO.

The revenue trajectory and the distillation clampdown are the two developments that matter for practitioners. If Anthropic is actively reducing distillation-driven usage, that changes the economics for labs that have been building on top of its outputs, and it may push those labs toward alternative model providers or toward training on their own data.

Anthropic quarterly revenue — \$bn

Q1		4.8
Q2		11.5
Q3		16.8

Revenue reported to investors for Q1, Q2 and Q3

KEY FACTS

Q1 REVENUE

about \$4.8bn

Q3 REVENUE

\$16.8bn

ADJUSTED OPERATING MARGIN (EXCL. SBC)

11%

Q2 REVENUE

\$11.5bn

Q3 REVENUE PER \$1 COMPUTE SPEND

\$1.70

DISTILLATION SHARE OF REVENUE

5% to 7%

WHY IT MATTERS

Anthropic is actively cutting back revenue from rival labs distilling its models, which means teams relying on that practice should expect tighter enforcement and may need to rework their training pipelines or switch providers.

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

09

HIGH IMPACT

Ars Technica

Targeted Ads Delivered 0-Click Mercenary Spyware to Al Jazeera's Devices

Mercenary spyware has been confirmed delivered through targeted ads with zero-click exploits, infecting 45 devices at Al Jazeera across 2024 and 2025.

The University of Toronto's Citizen Lab published findings Thursday naming the campaign AdFection and attributing the ad supply to Discover-Tech, an Israeli company presenting itself as a legitimate advertising firm. Citizen Lab describes Discover-Tech's primary purpose as bridging ad vendors and the mercenary spyware ecosystem, brokering bids that place ads on devices matching particular demographics. The exploits worked on both iOS and Android and required no interaction from the target: loading a website or opening an installed app was sufficient. Infection is invisible without forensic hardware or software.

The brokering mechanism relies on unique mobile IDs, location, IP addresses, and other breadcrumbs to assemble detailed profiles of millions of users. Citizen Lab obtained more than 4 billion bidding records handled by Discover-Tech. Those records show Discover-Tech ads were displayed in over 70,000 mobile apps and more than 120,000 websites between March and June 2026.

Citizen Lab frames the finding as likely the first confirmation of zero-click mercenary spyware delivered through targeted ads. The report warns that news organizations, individual journalists, editors, and other staff may be actively targeted through this vector, and that the distribution method poses a security risk to states and civil society at large, not only to direct victims.

KEY FACTS

DEVICES INFECTED AT AL JAZEERA

45

MOBILE APPS SHOWING DISCOVER-TECH ADS

70,000+

OBSERVATION WINDOW

March-June 2026

BIDDING RECORDS OBTAINED

4 billion+

WEBSITES SHOWING DISCOVER-TECH ADS

120,000+

PLATFORMS AFFECTED

iOS and Android

WHY IT MATTERS

Ad networks are now a confirmed delivery channel for zero-click mercenary malware, so any app or site embedding third-party ad SDKs is a potential infection surface for its users. Teams operating high-risk environments should treat ad traffic as an active threat vector and consider ad blocking, network filtering, and forensic device checks for at-risk staff.

[Read the original at Ars Technica](#) →

10

MEDIUM IMPACT

TechCrunch

AI Agents Ride Into Rivian's Cars — Anthropic and OpenAI Models Are Next

Rivian is preparing to let third-party AI agents from frontier labs run inside its vehicles, with CEO RJ Scaringe confirming at TechCrunch Disrupt that Anthropic, OpenAI and Meta models are on the roadmap.

Scaringe said Rivian has built its own in-house agentic framework that interprets a spoken request and routes it to the appropriate model — a local model running on the edge for simple commands, or a cloud model for more complex queries. He described the routing as "like a traffic light": the vehicle decides which model to use based on the question. As of May this year, the agent could already access external apps such as Google Calendar.

The next step, Scaringe said, is integrating products from frontier AI companies — he named Anthropic, and the article references OpenAI's Dots and Meta's Muse as plausible candidates. He did not give a timeline, and Rivian declined to elaborate, pointing to its "AI & Autonomy Day" event in December. Scaringe was explicit that vehicle-level controls are off-limits: "You wouldn't want an agent to determine, 'hey, it's time to apply the brakes, or time to steer the vehicle.'"

The integration is not trivial. Scaringe described "a lot of plumbing behind the scenes" and coordination with the labs themselves to enable it. The envisioned use case extends existing agentic app behavior into the car: a driver could ask the Rivian assistant to turn on ventilated seats, then in the same breath ask for help planning a dinner for guests on Friday, with the car pulling conversational context from an assistant interaction the driver had on another device.

Rivian is not first to this. Tesla added a feature earlier this year letting users talk to SpaceX's Grok chatbot and its Grok Bot assistant. Rivian's differentiator is its ground-up software stack — the company built its in-car software without Apple CarPlay or Android Auto and develops its own applications, including navigation.

KEY FACTS

THIRD-PARTY AGENTS NAMED

Anthropic (Claude), OpenAI Dots, Meta Muse

COMPANY EVENT

AI & Autonomy Day in December

EXTERNAL APP ACCESS

Google Calendar as of May

TESLA PRECEDENT

Grok chatbot and Grok Bot assistant added earlier this year

WHY IT MATTERS

For developers building agentic systems, Rivian's architecture is a concrete example of tiered model routing — edge models for low-stakes commands, cloud models for complex reasoning — with hard boundaries around safety-critical controls. The December AI & Autonomy Day will likely clarify which third-party models get first-class integration and what the developer surface looks like.

[Read the original at TechCrunch →](#)

11

MEDIUM IMPACT

TechCrunch

Amazon's Bee Doubles Down: Two All-Day AI Note-Taking Wearables From \$75

Amazon's Bee is shipping two new AI note-taking wearables that undercut Plaud on price while adding Amazon.com and Alexa+ integrations.

Bee, the personal AI wearable maker Amazon acquired last year, announced its first new products on Thursday: the Bee Frontier, a wrist- or clothing-worn device at \$74.99, and the Bee Eclipse, a credit-card-sized device that magnetically attaches to a smartphone at \$99.99. The Eclipse puts Bee in direct competition with the Plaud Note (\$159, or \$189 for the Note Pro) and the Pocket AI device (\$129), but at a lower price point. The Frontier ships in Onyx, Lavender, and Sage; the Eclipse in Silver, Midnight, Honey, and Blush.

Both devices capture everyday conversations and convert them into summaries, to-dos, and insights. Free use includes conversation summaries and recaps plus 14 days of memory. A \$11.99 per month Bee Pro subscription adds unlimited summaries, integrations with Amazon.com, Alexa+, and third-party apps, suggested to-dos, daily memories, a personal chat assistant, and conversation templates. The Frontier includes IP67 water resistance, up to seven days of battery life, USB-C charging, and dual microphones with advanced noise filtering.

On the security side, Bee uses a private computing architecture that does not store raw audio: audio is processed in real time and deleted, while transcripts are encrypted with keys the device creates and controls. Amazon employees cannot access those files unless the user manually shares them. Users can optionally connect Bee to Alexa so Echo devices use Voice ID to recognize the speaker and record conversations for insights; Amazon notes transcripts and summaries can be deleted at any time.

Bee also integrates with Gmail, Google Calendar, and Microsoft Outlook for drafting emails, setting meetings, and surfacing action items, and can accept photo uploads or connect to Apple HealthKit to fold wellness data into insights. The Frontier is available now on Amazon; the Eclipse will follow shortly. A redesigned mobile app and a new desktop app accompany both devices.

AI note-taking device prices — \$

Bee Frontier		74.99
Bee Eclipse		99.99
Pocket AI		129
Plaud Note		159
Plaud Note Pro		189

List prices of Bee and competing AI note-taking devices

KEY FACTS

BEE FRONTIER PRICE

\$74.99

PLAUD NOTE PRICE

\$159

POCKET AI PRICE

\$129

FREE MEMORY

14 days

BEE ECLIPSE PRICE

\$99.99

PLAUD NOTE PRO PRICE

\$189

BEE PRO SUBSCRIPTION

\$11.99/month

FRONTIER BATTERY LIFE

up to 7 days

WHY IT MATTERS

The \$74.99 entry price and encrypted, device-controlled transcript keys make always-on conversation capture cheaper and more privacy-defensible than existing Plaud-class devices, which could push more builders to treat ambient audio as a default input.

[Read the original at TechCrunch →](#)

12 **MEDIUM IMPACT** Google DeepMind

DeepMind Sends Antigravity to the Lab Bench — a New Antibiotic Candidate Out the Other Side

Google DeepMind has turned its agentic science platform Antigravity on antimicrobial discovery and produced a novel peptide candidate now headed to wet-lab validation.

The work centers on Google Antigravity, an agentic development platform that gives biologists a single workspace for discovering, structurally evaluating, and prioritizing protein targets. In the published walkthrough, researchers used Antigravity to search for antimicrobial peptides (AMPs) in *Streptococcus pseudopneumoniae*, setting a focused goal and steering the agent through an interactive, multi-turn session while it handled cross-database orchestration, structural analysis, and literature synthesis. The session surfaced a lead with predicted activity against pathogens involved in respiratory tract infections; DeepMind states it is now undergoing wet-lab validation.

A key component is ProtNLM (Protein Natural Language Model), a deep learning model trained to predict biological function directly from amino acid sequences. ProtNLM is used to enrich millions of previously uncharacterized UniProt entries with high-confidence functional labels, addressing the "dark matter" problem:

UniProt holds millions of sequences transferred from genome sequencing projects whose function is unknown because they lack similarity to known proteins. The platform also introduces the Biomolecular Structure Viewer, a 3D visualization tool embedded directly into Antigravity. The viewer is bidirectional: the agent can retrieve structures from the AlphaFold Database and highlight functional regions to explain hypotheses, while a researcher can mark residues or pin notes onto a binding pocket in 3D space, and those actions are synchronized back to the agent for reasoning over the exact structural context.

The framing is explicitly collaborative rather than autonomous. The researcher decides which questions to ask and what evidence to collect; the agent handles the heavy lifting. DeepMind's claim is not a new antibiotic, but a workflow compression: what previously required custom scripts for database queries, manual export to 3D modeling tools, disconnected genomic software, and coordination across experts for wet-lab assay design now happens in one session. The candidate itself is described as biologically plausible and backed by structural and genomic evidence, but explicitly still requires wet-lab validation.

What is genuinely new here is the integration: ProtNLM and AlphaFold-derived structures existed before, but Antigravity binds them into a shared, interactive loop with a 3D canvas that feeds researcher annotations back to the agent. The walkthrough is narrow — one amino acid window in one organism — but the same session structure can be redirected to different filtering criteria, sequence lengths, or other priority pathogens.

KEY FACTS

PROJECTED ANNUAL DEATHS FROM ANTIMICROBIAL RESISTANCE BY 2050

10 million

TARGET ORGANISM IN WALKTHROUGH

Streptococcus pseudopneumoniae

3D VISUALIZATION TOOL

Biomolecular Structure Viewer

LEAD CANDIDATE STATUS

Undergoing wet-lab validation

FUNCTION PREDICTION MODEL

ProtNLM

WHY IT MATTERS

For teams doing protein or peptide discovery, this is a concrete template for compressing multi-day database triage and structural screening into a single agent-assisted session, with the 3D viewer providing a mechanism for human structural intuition to feed back into the agent's reasoning. The wet-lab validation caveat is the boundary: the output is a prioritized lead, not a proven therapeutic.

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

Sources

- 01 Meta Ships Muse Spark Open: an 816B-Parameter Open-Weight Model Built for Macs
<https://techcrunch.com/2026/10/15/meta-launches-an-open-weight-model-that-runs-on-macs/>
- 02 Anthropic Opens the Door: METR Gets Employee-Level Access — and the Company Can't Veto Its Reports
<https://techcrunch.com/2026/10/15/anthropic-is-letting-metr-researchers-look-under-the-hood-with-employee-level-access/>
- 03 Scaling Indic Pretraining: 2.8 Trillion Tokens Through an Open Data Pipeline
<https://huggingface.co/blog/nvidia/scaling-indic-pretraining>
- 04 LLMs Prove They Can Design State-of-the-Art Quantum Circuits
<https://arxiv.org/abs/2610.22367>
- 05 Bluesky's Attie Turns Vibe Coding Into a Group Project — Sites and Games on the AT Protocol
<https://techcrunch.com/2026/10/15/bluesky-expands-its-ai-powered-attie-app-to-let-anyone-build-websites-games-and-more/>
- 06 FLUX.3 Image Klein Lands in Diffusers: 5.6B Open Weights With Editing and Layout Control
<https://huggingface.co/blog/flux-3>
- 07 Choosing a Chunking Strategy for Your RAG Stack: a Decision Tree With Failure Signals
<https://machinelearningmastery.com/choosing-a-chunking-strategy-for-your-llm-apps-a-decision-tree-approach/>
- 08 Anthropic Told Investors: \$16.8B Quarterly Revenue, Up Nearly 50%, IPO in Sight
<https://simonwillison.net/2026/Oct/15/anthropic-tells-investors/>
- 09 Targeted Ads Delivered 0-Click Mercenary Spyware to Al Jazeera's Devices
<https://arstechnica.com/security/2026/10/targeted-ads-delivered-0-click-exploits-to-install-mercenary-malware/>
- 10 AI Agents Ride Into Rivian's Cars — Anthropic and OpenAI Models Are Next
<https://techcrunch.com/2026/10/15/ai-agents-are-coming-to-rivians-cars/>
- 11 Amazon's Bee Doubles Down: Two All-Day AI Note-Taking Wearables From \$75
<https://techcrunch.com/2026/10/15/amazons-bee-launches-two-new-personal-ai-devices-starting-at-75/>
- 12 DeepMind Sends Antigravity to the Lab Bench — a New Antibiotic Candidate Out the Other Side
<https://deepmind.google/blog/accelerating-antimicrobial-discovery/>

About this document. Every story in the 16 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 2026-10-16T05:33:54Z · model deepseek-v4-pro:cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.