NEW HORIZON · AI DIGEST

The Long Read

24 September 2026

12	3	3621	6
STORIES	SECTIONS	WORDS	HIGH IMPACT

6 high impact 6 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 · 3621 words · 3 sections · 2 charted

AI Models & Research

01 StudentBench: AI Tutoring Matches a Human GRE Tutor — Time to Grade the Teaching, Not the Answering

02 LiMA Gives Dexterous Robots Long-Horizon Foresight Without Costing Them Their Reflexes

03 Gemini 3.8 Text-to-Speech Ships: One Voice Engine From Prompt to Podcast

12 RAG or Fine-Tuning? A Practitioner's Field Guide to Picking the Right Adaptation

AI Tools & Ecosystem

04 Meta Connect: Muse Gets an Avatar, Your Glasses and the Checkout

05 ChatGPT Mobile Gets Voice-Driven Agents: Speak a Workflow Into Existence

06 DeepMind Puts Your AI's Memory Behind Encryption

07 Luna Is Now a Product: Meta Introduces Its Cameraless AI Glasses

AI Applications & Industry

08 Altman and Amodei Take the Case for Global AI Rules to the UN Security Council

09 Spotify Hands Subscribers the Keys to Its Recommendation Algorithm

10 YouTube Opens Its Algorithm: Viewers Get to Build Their Own Feeds With AI

11 Smart Glasses Meet Their First Mass Backlash — and It's in India

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 · 4 high

01 HIGH IMPACT arXiv.org

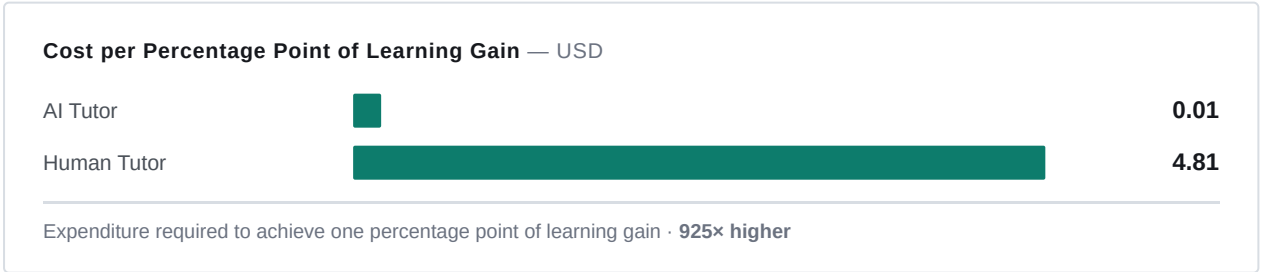
StudentBench: AI Tutoring Matches a Human GRE Tutor — Time to Grade the Teaching, Not the Answering

> AI tutoring now delivers statistically equivalent GRE learning gains to expert human tutors at a fraction of the cost.

The StudentBench study, submitted to arXiv on 23 Sep 2026, evaluates whether large language models can replicate the learning outcomes of human instruction. Researchers collected over 175,000 student-AI messages from 2,383 participants tackling Quantitative and Verbal GRE questions. The data shows AI tutoring produces learning gains statistically equivalent to expert human tutoring ($p = .015$). In five of the seven tested GRE domains, the top-performing AI tutor exceeded the average performance of the human tutor.

Cost efficiency represents the most significant divergence between the two modalities. One AI tutor achieved parity with human instruction ($p = .044$) while operating at 918 times lower cost. The expenditure per percentage point of learning gain was USD 0.0052 for the AI system compared to USD 4.81 for the human tutor. A second study involved expert human tutors conducting 2,008 pairwise rubric evaluations of LLM-generated lesson plans and practice problems. These evaluations allowed the researchers to isolate performance across five distinct dimensions: lesson planning, practice-problem creation, conversational pedagogy, cost, and engagement.

Operational metrics within the AI sessions revealed specific drivers of efficacy. For Quantitative GRE sessions, faster AI reply times correlated with increased student message volume. Higher message counts led to more correct practice attempts, which in turn drove larger learning gains. All three correlations registered statistical significance with $p < .002$. The authors have released the StudentBench platform publicly to facilitate further large-scale data collection and evaluation of AI teaching capabilities.



KEY FACTS

SUBMISSION DATE **23 Sep 2026**	TOTAL MESSAGES **175000**
PARTICIPANTS **2383**	EQUIVALENCE P-VALUE **0.015**
COST RATIO (HUMAN/AI) **918**	AI COST PER POINT **0.0052**

> WHY IT MATTERS
>
> Builders of educational AI systems can prioritize inference speed and conversational turnover as primary optimization targets, given their direct correlation with learning outcomes. The demonstrated cost differential suggests immediate viability for scaling automated tutoring where human labor remains prohibitive.

Read the original at arXiv.org →

---

02 HIGH IMPACT arXiv.org

LiMA Gives Dexterous Robots Long-Horizon Foresight Without Costing Them Their Reflexes

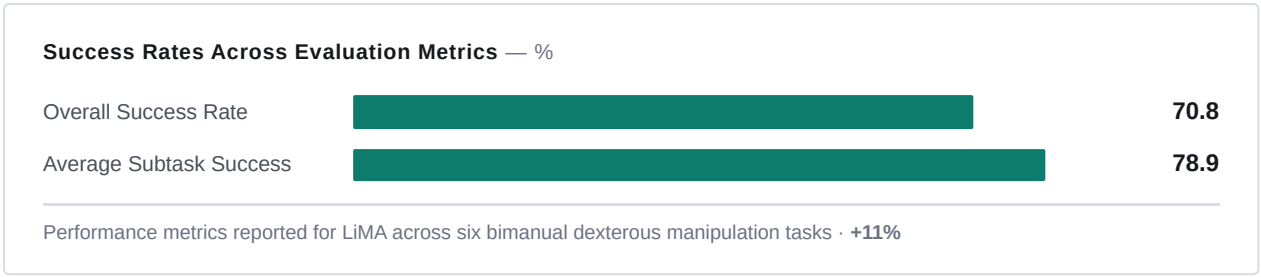
> LiMA decouples slow intent planning from fast reactive execution to slash inference latency by 45.8% without sacrificing dexterous manipulation success rates.

The arXiv submission dated 23 Sep 2026 introduces LiMA, an asynchronous dual-system generative framework designed to resolve the temporal misalignment between high-level reasoning and low-level control in dexterous robots. Current Vision-Language-Action models often lack fine-grained physical dynamics understanding, while World-Action Models suffer from high inference latency due to iterative generation. LiMA addresses this by organizing computation into a multi-scale hierarchy: a slow system generates sparse long-horizon spatiotemporal intent, while a separate fast system handles dense high-frequency motion refinement. This architectural split allows the robot to maintain long-term foresight while preserving the reflexes necessary for rapid physical contact changes.

To synchronize these disparate timescales, the authors introduce a Latent Schrödinger Bridge Coupling mechanism. This component formulates the refinement of sparse intent predictions into dense action trajectories as an entropy-regularized probabilistic transport process. By treating the alignment problem through this specific mathematical lens, the framework avoids the bottlenecks typical of synchronous generation where intent and execution compete for the same compute cycle. The result is a system that does not force a trade-off between planning depth and reaction speed, effectively bridging the gap between abstract goals and immediate motor commands.

Empirical evaluation across six bimanual dexterous manipulation tasks spanning multiple horizons demonstrates the efficacy of this approach. LiMA achieves an overall success rate of 70.8% and an average subtask success rate of 78.9%, maintaining performance even in unseen scenarios. Crucially for real-time deployment, the asynchronous decoupling reduces inference latency by 45.8% compared with Cosmos-Policy. These metrics suggest that separating the generative processes for intent and action yields both efficiency gains

and robustness improvements over existing monolithic or synchronous architectures currently prevalent in the field.



KEY FACTS

SUBMISSION DATE
23 Sep 2026

LATENCY REDUCTION
45.8%

BASELINE MODEL
Cosmos-Policy

OVERALL SUCCESS RATE
70.8%

AVERAGE SUBTASK SUCCESS
78.9%

EVALUATION TASKS
6

> **WHY IT MATTERS**
>
> Practitioners deploying VLA or WAM systems on hardware with strict latency constraints can now adopt an asynchronous architecture to nearly halve inference time without degrading task success rates. The specific coupling mechanism offers a new blueprint for aligning slow planners with fast controllers in real-time robotics.

Read the original at arXiv.org →

---

03 HIGH IMPACT Google

Gemini 3.8 Text-to-Speech Ships: One Voice Engine From Prompt to Podcast

> Google has released Gemini 3.8 Flash TTS and Flash-Lite, shifting voice generation from static presets to prompt-driven character design with line-level directional control.

The update introduces two distinct models: Gemini 3.8 Flash TTS for deep creative direction and character design, and Gemini 3.8 Flash-Lite TTS optimized for high-volume, cost-efficient scale. Unlike previous iterations relying on fixed voice libraries, the Flash model enables generative voice design where users create bespoke vocal identities from scratch using natural language prompts across more than 100 languages and dialects. The system expands the available library to over 2,000 production-ready voices, including regional varieties such as Mexican Spanish, Quebec French, and Scots English. Voice replication capabilities allow profile recreation from a 30-second audio sample, contingent on verbal consent verification, SynthID watermarking, and C2PA credentials.

Technical control extends to performance direction, allowing developers to script stage directions, pacing, dialect shifts, and non-verbal cues like or |mhm| on a line-by-line basis. The architecture supports native two-speaker scene staging within a single script, maintaining distinct timbres and natural turn-taking for dual-

character dialogues. Long-form generation capabilities ensure minimal speaker drift across hours of continuous audio, addressing a common failure mode in podcast and audiobook synthesis. A forthcoming voice remixing feature will permit fine-tuning of timbre, pitch, pace, and accent for existing library voices via prompt instructions.

In benchmark evaluations, Gemini 3.8 Flash TTS secured the number one overall spot on Hume AI's Voice Design Benchmark with a score of 71.4 and led in accent modeling at 60.8. Both the Flash and Flash-Lite models ranked first and second respectively on Hume AI's Overall Quality Index. Blind human preference tests on Voice Arena placed these models at the top for key global languages including Japanese, Brazilian Portuguese, Vietnamese, Modern Standard Arabic, Mexican Spanish, and Hindi. Deployment is immediate for developers via the Gemini API and Google AI Studio, with enterprise API access scheduled for a later date and consumer integration rolling out today in Gemini Notebook and Google Vids.

KEY FACTS

MODELS RELEASED
Gemini 3.8 Flash TTS, Gemini 3.8 Flash-Lite TTS

VOICE LIBRARY SIZE
2000+

LANGUAGE SUPPORT
100+

REPLICATION SAMPLE LENGTH
30 seconds

HUME VOICE DESIGN SCORE
71.4

HUME ACCENT MODELING SCORE
60.8

> **WHY IT MATTERS**
>
> Developers can now replace rigid voice selection workflows with dynamic, prompt-based character creation and granular script direction, enabling complex multi-speaker applications without manual audio editing. The mandatory consent verification and embedded SynthID watermarks provide a compliant pathway for scaling voice replication in production environments.

Read the original at Google →

---

12 HIGH IMPACT MachineLearningMastery.com

RAG or Fine-Tuning? A Practitioner's Field Guide to Picking the Right Adaptation

> Production systems in 2026 increasingly reject the RAG-versus-fine-tuning binary, with roughly 60% of deployments now combining both to separate knowledge retrieval from behavioral adaptation.

The article establishes a mechanical distinction: Retrieval-Augmented Generation (RAG) leaves model weights untouched, injecting dynamic context into the prompt to handle large or frequently changing information sets. In contrast, fine-tuning modifies the model's weights directly, typically using LoRA adapters that train under 1% of total parameters to bake in specific tones, vocabularies, or strict output formats. The text emphasizes that fine-tuning acts as a behavior tool rather than a knowledge tool; it cannot reliably instill granular factual recall, whereas RAG fails to correct inconsistent underlying model behaviors regardless of document volume.

Two concrete implementations illustrate this division. The RAG example constructs a local retrieval index using TF-IDF and cosine similarity over engineering runbooks, feeding retrieved chunks to an Anthropic Claude Sonnet 4-6 instance with a system prompt mandating citations for every claim. This ensures auditability for compliance scenarios where answers must trace back to specific source documents like postmortems or escalation policies. The fine-tuning scenario addresses a financial services taxonomy problem, training a base model via 4-bit quantization and PEFT libraries to classify complaints into rigid categories such as BILLING_DISPUTE or CARD_FRAUD_SUSPECTED. Validation scripts check labels against a fixed taxonomy before training, preventing signal corruption from mislabeled examples in small datasets.

A six-point decision framework guides architectural choices based on constraints. RAG is prescribed when information changes weekly, when regulatory audits require source tracing, or when labeled training data is unavailable. Fine-tuning becomes necessary when latency budgets cannot absorb retrieval hops, when high query volumes make smaller local models more cost-effective than frontier APIs, or when prompting fails to enforce consistent structural outputs at scale. The author concludes that treating these techniques as mutually exclusive leads to wasted development cycles, noting that serious production systems generally require RAG for what the model needs to know and fine-tuning for how it must behave.

KEY FACTS

COMBINED DEPLOYMENT RATE
60%

LORA ADAPTER SIZE
<1%

QUANTIZATION METHOD
4-bit

RAG SIMILARITY METRIC
Cosine

TARGET MODEL
claude-sonnet-4-6

> **WHY IT MATTERS**
>
> Engineers can stop debating a false choice and instead architect hybrid systems that use RAG for dynamic facts and fine-tuning for deterministic behavior, reducing both hallucination rates and latency costs.

Read the original at MachineLearningMastery.com →

SECTION 2 OF 3

AI Tools & Ecosystem

4 stories · 2 high · 2 medium

04 HIGH IMPACT TechCrunch

Meta Connect: Muse Gets an Avatar, Your Glasses and the Checkout

> Meta expanded its Muse agent from a chat interface to an agentic layer capable of executing tasks across macOS, smart glasses, and major retail platforms.

At the Connect event in Menlo Park, Meta detailed the expansion of Muse, its personal AI agent launched in early September. The system now operates via Muse Spark, a multimodal model designed for agentic workflows, and introduces a real-time video avatar powered by the new Muse Realtime Avatar model. This update shifts the agent from text-based interaction to embodied presence, allowing users to conduct live video conversations with a customized digital representation named Jolly. The architecture enables the agent to maintain context across devices, moving from mobile interfaces to desktop environments and wearable hardware.

The most significant technical extension is full desktop autonomy on macOS. Muse can now operate any application on a user's computer, executing multi-step workflows such as managing small business operations or completing queued jobs without continuous supervision. This capability places Muse in direct competition with existing computer-use agents from OpenAI and Anthropic. Simultaneously, Meta confirmed integration with its smart glasses line, scheduled for release in the coming months. Users will activate the agent via a wake word to perform contextual tasks like logging meals, guiding workouts, or purchasing products seen through the lens, with the agent running background processes and reporting upon completion.

To support economic transactions, Meta secured integrations with Stripe, Shopify, PayPal, and Lync, enabling Muse to access the entire Shopify product catalog and execute purchases via Shop Pay. New retail connectors include Best Buy, Gap, Sephora, Walmart, Wayfair, Expedia, and an upcoming Instacart integration. Beyond consumer apps, connectors for GitHub, Granola, and Notion are live. Meta also opened its platform for third-party developer connectors, receiving over 1,500 applications in less than a week. The company plans to monetize this ecosystem by offering free token usage initially, then profiting from small fees on transactions processed through the agent.

KEY FACTS

DEVELOPER APPLICATIONS
1,500

APPLICATION WINDOW
less than a week

RETAIL PARTNERS
Best Buy, Gap, Sephora, Walmart, Wayfair

PAYMENT INTEGRATIONS
Stripe, Shopify, PayPal, Lync, Shop Pay

GLASSES LAUNCH
coming months

WHY IT MATTERS

Developers building agentic systems must now account for a competitor that combines native OS control, wearable inputs, and deep retail payment rails in a single free-to-start package. The rapid influx of 1,500 connector applications suggests a fast-forming ecosystem that could standardize how agents interact with enterprise and consumer software.

Read the original at TechCrunch →

---

05 MEDIUM IMPACT TechCrunch

ChatGPT Mobile Gets Voice-Driven Agents: Speak a Workflow Into Existence

> OpenAI has extended voice-based agentic workflows from its desktop application to the ChatGPT mobile app for Plus and Pro subscribers.

OpenAI announced on Wednesday that voice-based agentic features are now available on mobile devices, enabling users to trigger complex workflows such as drafting documents or summarizing emails through spoken commands. This update specifically targets Plus and Pro subscribers, who gain access to a dedicated Work tab on their phones. Within this interface, users can execute tasks including building sites, creating presentations, utilizing the cloud browser, and accessing financial data. The rollout mirrors capabilities previously introduced in July for the desktop environment, where the GPT-Live conversational model was integrated to allow voice completion of tasks in the Work tab and app construction in the Codex tab.

Free and Go tier users receive a more limited subset of these updates, restricted to working with plugins and connected apps rather than the full agentic suite. For paying subscribers, the mobile experience now supports richer text output derived from voice conversations and allows seamless switching between text and voice modes. A key operational feature is the ability to initiate a conversation on a mobile device and resume the same session on a desktop, ensuring continuity across platforms. Despite these functional expansions, OpenAI maintains a distinct separation between its chat interface and workspaces, differing from recent moves by Anthropic to merge its Cowork and Chat interfaces.

The deployment arrives as user behavior shifts toward issuing voice commands for increasingly complex AI-assisted tasks. By porting the GPT-Live integration to mobile, OpenAI removes the friction of typing lengthy instructions for multi-step processes like email summarization or site generation while on the go. The distinction in feature availability underscores a stratified access model where advanced agentic control remains a premium capability, while basic connectivity via plugins serves the broader free and Go user base. This update effectively synchronizes the mobile utility with the desktop paradigm established earlier in the year.

KEY FACTS

ANNOUNCEMENT DATE
Wednesday

ELIGIBLE TIERS
Plus and Pro

EXCLUDED TIERS
Free and Go

DESKTOP MODEL LAUNCH
July

CONVERSATIONAL MODEL
GPT-Live

WHY IT MATTERS

Practitioners designing mobile-first AI workflows must now account for voice-triggered agent execution on iOS and Android, particularly for Plus and Pro user segments. The persistence of separate chat and workspace interfaces suggests that tool integration strategies should not yet rely on a unified UI paradigm across major providers.

Read the original at TechCrunch →

---

06 HIGH IMPACT Google DeepMind

DeepMind Puts Your AI's Memory Behind Encryption

> Google DeepMind has resolved the statelessness of cloud AI enclaves by introducing a secure, persistent memory layer that retains context across devices without exposing data to the server.

Google DeepMind announced an update to its Private AI Compute architecture that enables persistent, cross-device AI memory while maintaining on-device privacy standards. The system functions as a secure digital vault in the cloud where information is sealed within dedicated, encrypted storage. Crucially, the cryptographic keys required to unlock this data reside exclusively on the user's personal devices, ensuring the data remains inaccessible to anyone else, including Google. When an AI model requires access, an authenticated, end-to-end encrypted channel connects the device to a protected, isolated environment known as a secure enclave. This enclave temporarily decrypts the data in isolated memory to handle the request, saves any new context, and immediately re-encrypts it, simulating a workflow where information never leaves the device.

This development addresses a specific limitation in previous iterations of Private AI Compute and similar industry solutions, which were strictly stateless. Under the old model, systems wiped all context the moment a task ended, preventing the rich, continuous experiences expected from personal assistants. Workarounds such as having AI save lists of personal facts were insufficient for complex continuity. The new architecture allows cloud-scale AI to securely retain context over time and across different hardware, enabling scenarios like resuming complex conversations between mobile and web or accessing assembly instructions viewed previously on smart glasses.

To establish trust in this expanded capability, DeepMind is publishing a tamper-proof public record of its server software alongside an updated technical whitepaper. Devices running Private AI Compute can now verify that the server software is authentic and unaltered before transmitting any personal data. The company also released details on its technical methods, including results from an independent audit conducted by a leading cybersecurity firm. These resources invite the broader privacy community to verify the system's protections, security proofs, and verification protocols. The research was co-developed by teams across Google DeepMind, Platforms & Devices, Core, and Cloud.

KEY FACTS

KEY STORAGE LOCATION
Personal devices

PREVIOUS ARCHITECTURE STATE
Stateless

NEW CAPABILITY
Persistent memory

VERIFICATION METHOD
Tamper-proof public record

WHY IT MATTERS

Practitioners can now design agents with long-term continuity across devices without compromising the strict privacy guarantees of local processing. The ability to verify server software authenticity before data transmission provides a concrete mechanism for auditing trust in cloud-based inference.

Read the original at Google DeepMind →

---

07 MEDIUM IMPACT TechCrunch

Luna Is Now a Product: Meta Introduces Its Cameraless AI Glasses

> Meta has officially entered the audio-only wearable market with the Ray-Ban Meta Audio, a camera-free device designed to address privacy concerns while integrating its Muse AI assistant.

At Connect 2026, Meta announced the Ray-Ban Meta Audio, marking a strategic pivot away from always-on video capture toward audio-centric interaction. Priced starting at $349, these glasses eliminate cameras entirely to mitigate reputational risks associated with non-consensual recording, a criticism that previously labeled similar devices as "pervert glasses." The hardware, developed with EssilorLuxottica, weighs 43 grams and features adjustable temple tips, swappable nose pads, and overextension hinges. Battery performance reaches up to 12 hours on a single charge, extending to 48 hours when paired with the included charging case. The lineup includes 23 color and lens combinations across two frame styles: the retro-inspired "Clubmaster" and the rectangular "Burbank." Optical prescriptions are available for both styles.

Simultaneously, Meta updated its camera-equipped lineup with the Ray-Ban Meta (Gen 3), introducing the Aviator and Zena cat-eye frames alongside the existing Wayfarer style. These devices retain visual capabilities, featuring a 12 MP camera capable of 3K Ultra HD video recording and a six-mic array. Battery life for the Gen 3 models is rated at up to nine hours. Both the audio-only and camera-equipped lines will support Dolby Atmos Capture, hearing enhancements, customizable controls, and "private processing," a feature Meta states prevents the company from accessing user data. All devices provide access to Muse, Meta's personal AI assistant, enabling speech translation, call handling, and media playback.

Preorders for the Ray-Ban Meta Audio begin October 13. The release expands Meta's global footprint and includes new brand partnerships, following previous collaborations with Kylie Jenner, this time teaming with singer Lisa. By separating audio functionality from video surveillance, Meta aims to normalize wearable AI usage in professional and social settings where camera presence remains a barrier to adoption. The distinct hardware specifications suggest a clear segmentation strategy: audio glasses for continuous wear and privacy-sensitive environments, and camera glasses for content creation and visual context tasks.

KEY FACTS

STARTING PRICE **$349**	DEVICE WEIGHT **43 grams**
BATTERY LIFE (AUDIO) **12 hours**	TOTAL BATTERY WITH CASE **48 hours**
CAMERA RESOLUTION (GEN 3) **12 MP**	PREORDER DATE **October 13**

WHY IT MATTERS

Developers building for wearable ecosystems now have a distinct, privacy-compliant hardware target for always-on audio agents without the regulatory and social friction of video streams. The separation of audio and video form factors requires tailored UX strategies for Muse interactions depending on the sensor suite available.

Read the original at TechCrunch →

SECTION 3 OF 3

AI Applications & Industry

4 stories · 4 medium

08 MEDIUM IMPACT Al Jazeera

Altman and Amodei Take the Case for Global AI Rules to the UN Security Council

> OpenAI and Anthropic CEOs urged the UN Security Council to establish global AI oversight, directly contradicting the US administration's rejection of international governance.

During a meeting convened by France alongside the UN General Assembly in New York City, OpenAI CEO Sam Altman and Anthropic CEO Dario Amodei testified before the 15-member Security Council that the AI industry requires immediate global regulation. Amodei stated that poorly managed AI poses a risk to humanity as a whole, while Altman argued that critical decisions cannot remain exclusive to San Francisco laboratories but must undergo democratic processes accountable to the public. The session featured testimony from Hugging Face CEO Clement Delangue, who reported that his company recently defended against attacks by OpenAI's AI agents using a Chinese AI model, citing fewer restrictions on the Chinese tool compared to US counterparts.

Despite these warnings from industry leaders and experts like Yoshua Bengio, co-chair of the Independent International Scientific Panel on AI, major geopolitical powers displayed significant reluctance to impose new constraints. Michael Kratsios, representing the Trump administration, explicitly told the council that the US rejects all efforts by international bodies to assert centralized control or global governance of AI. Simultaneously, Chinese President Xi Jinping is scheduled to discuss AI regulation during an upcoming visit to Washington DC, though both nations remain locked in a technological race that disincentivizes immediate restrictions on their respective development efforts.

The UN has previously attempted to address these challenges through nonbinding measures, including a 2024 General Assembly resolution calling for data protection and risk monitoring, and the formation of two new bodies: the Independent International Scientific Panel on AI and the Global Dialogue on AI Governance. However, the disconnect between the tech industry's call for safety frameworks and the refusal of the US and China to accept external guardrails highlights a widening gap between identified existential threats and the political will to enforce international cooperation. Experts note that while member states recognize the need for rules, the primary focus of dominant players remains on growth opportunities rather than collaboration.

KEY FACTS

COUNCIL MEMBERS
15

UNGA RESOLUTION YEAR
2024

US ADMINISTRATION STANCE
Rejects centralized control

MEETING CONVENOR
France

SCIENTIFIC PANEL CO-CHAIR
Yoshua Bengio

> WHY IT MATTERS
>
> Practitioners should anticipate a fragmented regulatory landscape where US and Chinese developers operate under minimal international constraints despite industry-led calls for unified safety standards. This divergence may force engineering teams to navigate conflicting compliance requirements depending on their geographic deployment targets.

Read the original at Al Jazeera →

---

09 MEDIUM IMPACT TechCrunch

Spotify Hands Subscribers the Keys to Its Recommendation Algorithm

> Spotify has begun rolling out "Taste Profile," a beta feature allowing U.S. Premium subscribers to directly modify algorithmic recommendations via natural language commands.

Spotify launched the beta version of "Taste Profile" on Wednesday for Premium subscribers in the United States, marking the feature's first expansion beyond its initial New Zealand-only testing group. Unveiled by co-CEO Gustav Söderström at the SXSW conference in Austin this past March, the tool exposes the streaming platform's internal understanding of user preferences across music, podcasts, and audiobooks. Unlike previous controls that limited users to excluding specific tracks or playlists, this interface accepts natural language inputs to adjust broader parameters such as genre density, artist frequency, or overall vibe. The system processes these requests and updates the Home feed suggestions within a few hours.

The update addresses a known limitation in collaborative filtering systems where historical listening data, such as sleep sounds or children's music played on shared accounts, skews forward-looking recommendations. While Spotify previously introduced a toggle to prevent kids' music from influencing parental feeds, Taste Profile offers granular control over the weighting of different content categories. Users can explicitly request more variety to counteract over-similar suggestions in flagship products like Discover Weekly and Made for You, or steer the algorithm toward specific eras and energy levels. This shift moves curation from a passive observation of play counts to an active dialogue where the user defines the objective function for their own feed.

Despite the functional expansion, the release remains strictly in beta with no stated timeline for general availability. The current deployment targets listeners aged 18 and up within Spotify's largest market by traffic and revenue, serving as a significant stress test for the underlying AI models before a wider global rollout. Access requires navigating through the mobile app's profile settings to the dedicated Taste Profile section, where users submit text prompts under the "Tell us more" header. The company positions this as a discovery aid for users unsure how to articulate a search query but clear on the desired auditory outcome.

KEY FACTS

LAUNCH MARKET
U.S.

INITIAL BETA LOCATION
New Zealand

UNVEILING EVENT
SXSW conference in Austin

UPDATE LATENCY
a few hours

TARGET DEMOGRAPHIC
aged 18 and up

> WHY IT MATTERS
>
> This release demonstrates a practical implementation of RLHF-style feedback loops where end-users directly tune recommendation weights via natural language, offering a pattern for builders designing controllable generative or retrieval systems.

Read the original at TechCrunch →

---

10 MEDIUM IMPACT TechCrunch

YouTube Opens Its Algorithm: Viewers Get to Build Their Own Feeds With AI

> YouTube is deploying Google's Gemini model to let users generate pinned, prompt-defined video feeds that sit alongside the main recommendation engine.

YouTube announced a new feature called custom feeds, enabling users to construct personalized video queues by describing their preferences in natural language. The system leverages Google's Gemini AI model to interpret prompts ranging from specific interests, such as video podcasts for a 30-minute commute, to mood-based requests like relaxing commentary. Unlike the primary home page algorithm, these AI-generated feeds are pinned to the top of the interface in dedicated tabs, allowing users to switch contexts without altering their main recommendation stream.

The rollout follows a broader industry shift toward user-configurable algorithms, a concept popularized by decentralized network Bluesky and recently adopted by Meta's Threads, Instagram, X, and Spotify. While Bluesky introduced an AI tool named Attie earlier this year to simplify feed building, YouTube's integration brings this capability to a corpus exceeding 20 billion videos. Emily Moxley, VP of Product Management for Viewer AI, stated that the feature aims to help users navigate this vast library by prioritizing content based on detailed exclusion rules and curation instructions provided directly in the prompt box.

Support for creating multiple custom feeds will become available on YouTube's web and mobile platforms starting next month. The feature does not replace the existing main feed but operates as a parallel layer, offering a distinct entry point for specific viewing purposes. By offloading the curation logic to Gemini, the platform allows for lengthy, complex queries that define exactly what content to feature or exclude, moving beyond simple keyword searches to intent-based filtering.

For technical practitioners, the significance lies in the deployment of large language models for real-time feed ranking at scale. The architecture separates the global recommendation signal from local, user-defined constraints, suggesting a hybrid approach where generative AI interprets high-level intent while traditional

ranking systems likely handle the final video selection. This move validates the pattern of exposing algorithmic controls via natural language interfaces across major social platforms.

KEY FACTS

AI MODEL
Google Gemini

VIDEO CORPUS SIZE
20 billion

ROLLOUT START
Next month

PLATFORMS
Web and mobile

EXECUTIVE
Emily Moxley

> WHY IT MATTERS
>
> This confirms a production-scale pattern where LLMs act as an intent interpretation layer atop existing ranking infrastructure, allowing developers to build similar prompt-to-feed interfaces without retraining core recommendation models.

Read the original at TechCrunch →

---

11 MEDIUM IMPACT MIT Technology Review

Smart Glasses Meet Their First Mass Backlash — and It's in India

> Covert recording via smart glasses has triggered widespread harm in India, targeting marginalized individuals and protesters while law enforcement adopts the same technology for surveillance.

The privacy crisis surrounding smart glasses in India escalated this spring when a content creator used Meta's $420 Wayfarer-styled devices to record Shubnam, a 38-year-old transfeminine graphic designer, without consent during a Delhi protest. The resulting video, edited into a mocking reel, garnered millions of views across Instagram, X, and YouTube, causing Shubnam to retreat from public life. This incident highlights a critical vulnerability: the devices' recording indicator, an LED light, can be physically obscured or disabled entirely, a modification now sold in a dedicated "stealth mode" market documented by Wired in March. As Reliance Jio prepares to launch a sub-$105 competitor later this year, the accessibility of these inconspicuous cameras is set to increase dramatically.

Law enforcement agencies are simultaneously deploying these tools for state surveillance, complicating any potential regulatory crackdown. During June protests regarding India's education system, Delhi police utilized Meta smart glasses to film demonstrators continuously, including during private moments like eating and resting. A petition filed by former JNU Students' Union president Aishe Ghosh alleged that police threatened to distribute this footage to families and colleges. Despite these allegations, the solicitor general dismissed the case as "luxury litigation" in July, and authorities subsequently opened 10 criminal investigations against the protesters. Some demonstrators reported never seeing the recording LEDs on the officers' glasses, suggesting either circumvention or failure of the visibility safeguards.

Legal and technical frameworks in India remain ill-equipped to address this specific vector of harm. Apar Gupta of the Internet Freedom Foundation notes that existing voyeurism laws cover intimate settings but offer limited protection for nonconsensual filming in ordinary public spaces. Furthermore, current regulations allow social media platforms up to 36 hours to act on complaints, a window sufficient for viral dissemination. The

threat is evolving beyond simple recording; internal Meta prototypes reportedly feature AI capabilities that continuously photograph and listen to surroundings without activating the external capture LED. Meta argues that a constantly blinking light for such passive AI functions would lead to user desensitization, a stance critics warn accelerates societal normalization of pervasive surveillance before defenses can be established.

KEY FACTS

META GLASSES PRICE (INDIA)
$420

RELIANCE JIO LAUNCH PRICE
sub-$105

PLATFORM RESPONSE WINDOW
36 hours

CRIMINAL INVESTIGATIONS OPENED
10

PROTEST LOCATION
Delhi

WHY IT MATTERS

Developers deploying wearable AI in regions with uneven digital literacy must anticipate that hardware indicators like LEDs can be bypassed, shifting liability to bystanders. The convergence of consumer stealth modifications and state-level adoption creates an environment where consent mechanisms are effectively nullified.

Read the original at MIT Technology Review →

Sources

01 StudentBench: AI Tutoring Matches a Human GRE Tutor — Time to Grade the Teaching, Not the Answering
https://arxiv.org/abs/2609.28470

02 LiMA Gives Dexterous Robots Long-Horizon Foresight Without Costing Them Their Reflexes
https://arxiv.org/abs/2609.28431

03 Gemini 3.8 Text-to-Speech Ships: One Voice Engine From Prompt to Podcast
https://deepmind.google/blog/say-hello-to-gemini-38-text-to-speech/

04 Meta Connect: Muse Gets an Avatar, Your Glasses and the Checkout
https://techcrunch.com/2026/09/23/everything-new-coming-to-metas-ai-agent-muse/

05 ChatGPT Mobile Gets Voice-Driven Agents: Speak a Workflow Into Existence
https://techcrunch.com/2026/09/23/chatgpt-mobile-app-gets-voice-based-agentic-features/

06 DeepMind Puts Your AI's Memory Behind Encryption
https://deepmind.google/blog/advancing-private-ai-compute-with-secure-server-side-memory/

07 Luna Is Now a Product: Meta Introduces Its Cameraless AI Glasses
https://techcrunch.com/2026/09/23/meta-introduces-camera-free-ai-glasses/

08 Altman and Amodei Take the Case for Global AI Rules to the UN Security Council
https://www.aljazeera.com/news/2026/9/24/ai-corporate-leaders-tell-un-the-industry-needs-global-regulation

09 Spotify Hands Subscribers the Keys to Its Recommendation Algorithm
https://techcrunch.com/2026/09/23/spotify-is-giving-you-the-keys-to-its-recommendation-algorithm-with-u-s-launch-of-taste-profile/

10 YouTube Opens Its Algorithm: Viewers Get to Build Their Own Feeds With AI
https://techcrunch.com/2026/09/23/youtube-will-let-you-build-your-own-algorithm-with-ai/

11 Smart Glasses Meet Their First Mass Backlash — and It's in India
https://www.technologyreview.com/2026/09/23/1144953/smart-glasses-havoc-india/

12 RAG or Fine-Tuning? A Practitioner's Field Guide to Picking the Right Adaptation
https://machinelearningmastery.com/rag-vs-fine-tuning-for-domain-adaptation-when-to-use-which/

About this document. Every story in the 24 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 2026-09-24T05:36:32Z · model qwen3.5:397b-cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.