Member-only story

The Week AI Took Over: 8 Releases That Changed Everything

From video generation with synchronized sound to instant shopping in ChatGPT, one extraordinary week redefined what artificial intelligence can do — and where it's heading next.

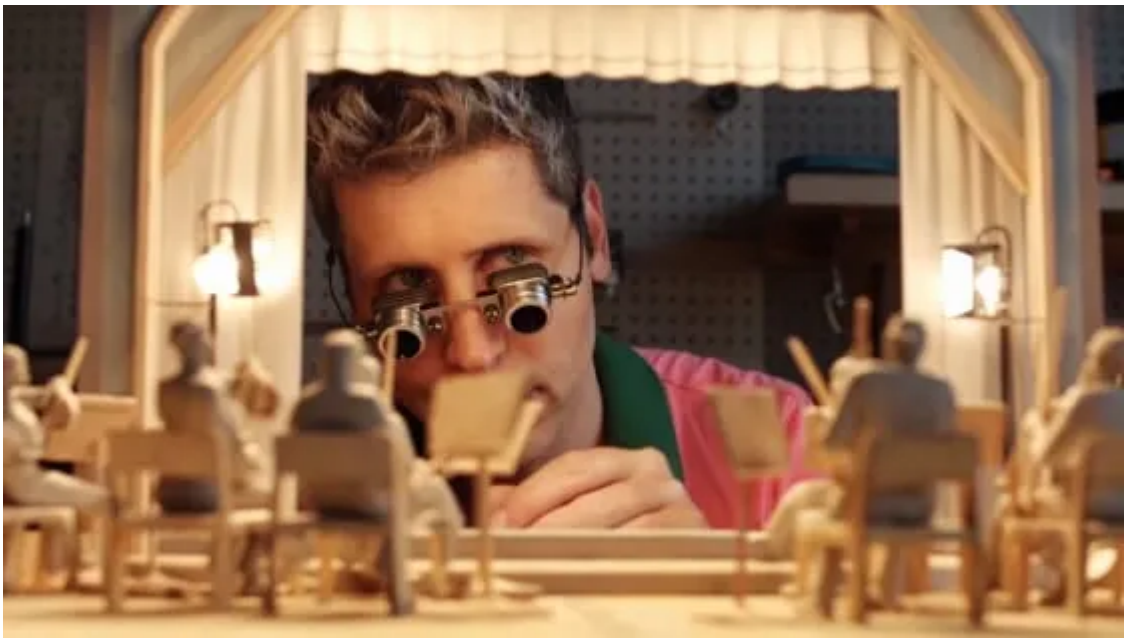
8 min read · Oct 6, 2025

Adham Khaled Following

Listen Share More

The last week of September 2025 will be remembered as one of the most consequential periods in artificial intelligence history. In just seven days, eight groundbreaking releases emerged from leading AI companies, each pushing boundaries in video generation, language models, coding agents, and agentic commerce. What we witnessed wasn't just incremental progress — it was a collective leap forward that signals AI's transition from experimental technology to production-ready systems reshaping industries.

OpenAI Sora 2: The GPT-3.5 Moment for Video



On September 30, 2025, OpenAI launched Sora 2, calling it “the GPT-3.5 moment for video” — and the comparison holds weight. Unlike its February 2024 predecessor, Sora 2 generates up to 20-second videos at 1080p with synchronized dialogue, sound effects, and realistic physics that obey the laws of nature. When a basketball player misses a shot in Sora 2, the ball rebounds off the backboard instead of teleporting to the hoop — a crucial distinction that demonstrates genuine world simulation.

The revolutionary Cameos feature allows users to insert themselves into AI-generated scenes with remarkable fidelity after a one-time video recording. OpenAI released a new iOS app with TikTok-style swipe navigation, positioning Sora as both a creative tool and social platform. Within 72 hours, viral videos flooded the app — from Sam Altman “stealing” GPUs at Target to Rick and Morty in Minecraft.

Technical capabilities: Sora 2 handles Olympic gymnastics routines, triple axels, and complex object interactions with unprecedented realism. The model excels at cinematic and anime styles while maintaining consistent world state across multiple shots. Currently available in the U.S. and Canada, the app operates on an invite-only basis with plans for API access.

Claude Sonnet 4.5: The World’s Best Coding Model

Anthropic raised the stakes on September 29, 2025, by releasing Claude Sonnet 4.5, which the company boldly claims is “the best coding model in the world”. The model achieves state-of-the-art performance on SWE-bench Verified, scoring 77.2% on real-world software coding tasks. More impressively, enterprise customers observed Claude Sonnet 4.5 coding autonomously for up to 30 hours during trials, building applications, purchasing domain names, and performing SOC 2 audits.

On OSWorld — a benchmark testing real-world computer tasks — Sonnet 4.5 leads at 61.4%, compared to its predecessor’s 42.2% just four months earlier. The model demonstrates substantial gains in reasoning, math, and domain-specific knowledge across finance, law, medicine, and STEM fields.

Developer adoption: Major platforms immediately integrated the model. Cursor CEO Michael Truell called it “state-of-the-art coding performance,” while Windsurf CEO Jeff Wang described it as representing “a new generation of coding models”. Pricing remains unchanged at $3 per million input tokens and $15 per million output tokens.

Anthropic also released the Claude Agent SDK, providing developers with the same infrastructure powering Claude Code, including memory management, permission systems, and subagent coordination.

DeepSeek-V3.2-Exp: Breakthrough in Sparse Attention

Chinese AI company DeepSeek released V3.2-Exp on September 29, 2025, introducing DeepSeek Sparse Attention (DSA) — the first implementation of fine-grained sparse attention in a production model. Built on the 671B-parameter V3.1-Terminus architecture, the experimental model dramatically improves long-context processing efficiency while maintaining comparable performance.

The economic impact is substantial: API pricing dropped by over 50%, with input costs as low as $0.07 per million tokens when utilizing cache hits. The sparse attention mechanism selectively computes attention weights, reducing computational complexity without sacrificing output quality.

Open-source commitment: DeepSeek released complete inference code, CUDA kernels, and a technical report documenting the DSA architecture. The company positions V3.2-Exp as an intermediate step toward V4, laying groundwork for next-generation architecture.

Z.ai GLM-4.6: China's Coding Agent Challenger

Beijing-based Z.ai (formerly Zhipu) unveiled GLM-4.6 on September 30, 2025, with enhanced coding and agentic capabilities designed to compete directly with Anthropic and OpenAI. Self-reported benchmarks across eight evaluations show improvements in coding, reasoning, and agency compared to July's GLM-4.5.

While GLM-4.6 demonstrates "competitive advantages" over Claude Sonnet 4 and DeepSeek-V3.2-Exp, it still trails Claude Sonnet 4.5 in pure coding ability. The release underscores China's strategic focus on coding models as a pathway toward artificial general intelligence.

Agentic capabilities: GLM-4.6 emphasizes advanced agentic features, positioning it for enterprise applications requiring autonomous task execution. The model supports open-source accessibility, broadening its potential adoption across research institutions and developers.

Thinking Machines Tinker: Fine-Tuning for the Masses

On October 1, 2025, Thinking Machines Lab — the $12 billion AI startup co-founded by former OpenAI CTO Mira Murati — launched Tinker, a Python-based API for

distributed LLM fine-tuning. Rather than releasing a foundational model, the company delivered a managed service that automates the complex process of customizing open-source models.

Technical features: Tinker provides low-level control through functions like `forward_backward` and `optim_step` while handling backend infrastructure including scheduling, resource management, and failure recovery on GPU clusters. The platform supports models from Meta's Llama 3.1 to large mixture-of-experts architectures like Qwen3-235B, utilizing Low-Rank Adaptation (LoRA) for efficient fine-tuning.

Currently in private beta, Tinker received praise from Princeton, Stanford, and Berkeley researchers who appreciated focusing on research rather than engineering overhead. The company released an open-source Tinker Cookbook documenting post-training techniques.

ChatGPT Instant Checkout: Agentic Commerce Arrives

OpenAI launched Instant Checkout on September 29, 2025, enabling U.S. users to purchase products directly within ChatGPT conversations. The feature, powered by the open-source Agentic Commerce Protocol co-developed with Stripe, fundamentally reshapes e-commerce by moving discovery and transactions inside conversational AI.

How it works: Users ask shopping questions like “waterproof hiking backpack under $200,” and ChatGPT surfaces relevant products from integrated merchants. Tapping “Buy” completes the purchase using Apple Pay, Google Pay, or credit card without leaving the conversation.

Merchant integration: The feature launched with Etsy and is expanding to over 1 million Shopify merchants including Glossier, Skims, Spanx, and Vuori. Product ranking depends on match quality, availability, and seller status — no paid ads or boosts. Merchants retain full operational ownership of fulfillment, returns, and customer service.

This positions OpenAI in direct competition with Google Search and Amazon for e-commerce traffic, potentially disrupting how consumers discover and purchase products online.

Google DeepMind Dreamer 4: Learning in Imagined Worlds

Google DeepMind introduced Dreamer 4 on September 28, 2025, a reinforcement learning agent that trains entirely inside its own world model. The breakthrough demonstrates unprecedented capabilities: Dreamer 4 became the first agent to obtain diamonds in Minecraft purely from offline data, without environment interaction.

Technical innovation: The task requires choosing sequences of over 20,000 mouse and keyboard actions from raw pixels. Dreamer 4 significantly outperforms OpenAI’s VPT offline agent while using 100 times less data. The world model achieves real-time interactive inference on a single GPU through a new objective and architecture.

Imagination training: The agent practices tasks inside its world model — what researchers call “imagination training” — allowing it to learn complex, long-horizon behaviors without costly real-world interactions. This capability proves particularly valuable for robotics applications where physical experimentation is expensive or impractical.

The AI Spending Report: Where Startup Dollars Go

Andreessen Horowitz released its first AI Application Spending Report on October 2, 2025, analyzing transaction data from Mercury's 200,000+ customers over June-August 2025. The report identifies the top 50 AI-native application layer companies startups actually pay for, revealing real-time patterns in AI adoption.

Key findings: OpenAI ranked first in spending, with Anthropic second. Vibe-coding tools dominated with Replit at #3, Cursor at #6, and Lovable at #18. The report shows proliferation rather than consolidation — businesses embrace multiple specialized AI tools rather than standardizing on one or two per category.

Investment context: AI startups attracted a record $192.7 billion in venture capital through 2025, representing more than half of global VC funding for the first time. U.S. investors dedicated 62.7% of total investments to AI-focused companies, while global investors allocated 53.2%.

What This Week Means for AI's Future

These eight releases collectively demonstrate AI's maturation from research prototypes to production systems. Video generation now matches language model capabilities from 2022. Coding agents autonomously handle 30-hour tasks. Fine-tuning becomes accessible to researchers without infrastructure expertise. Commerce moves inside conversational interfaces. World models enable offline learning at scale.

Industry implications: The rapid-fire releases reveal intensifying competition among AI labs racing toward artificial general intelligence. Coding models emerge as a critical battleground, with Anthropic, OpenAI, DeepSeek, and Z.ai all prioritizing software engineering capabilities. Meanwhile, startups face bifurcated funding landscapes — abundant capital flows to proven AI companies while early-stage ventures struggle.

The week also highlights AI's expanding surface area. Video generation enters mainstream creativity. Language models master computer use and agentic workflows. Infrastructure enables customization without expertise barriers. Commerce protocols integrate purchasing into natural language interactions.

As OpenAI's Sora team wrote: "Video models are getting very good, very quickly. General-purpose world simulators and robotic agents will fundamentally reshape society and accelerate the arc of human progress". Based on this week's releases, that future arrived faster than anyone anticipated.

A message from our Founder

Hey, Sunil here. I wanted to take a moment to thank you for reading until the end and for being a part of this community.

Did you know that our team run these publications as a volunteer effort to over 3.5m monthly readers? **We don't receive any funding, we do this to support the community.** ❤️

If you want to show some love, please take a moment to **follow me on LinkedIn, TikTok, Instagram**. You can also subscribe to our **weekly newsletter**.

And before you go, don't forget to **clap** and **follow** the writer!

Artificial Intelligence | Software Development | Technology | Coding | Data Science

Follow

Published in Artificial Intelligence in Plain English

32K followers · Last published just now

New AI, ML and Data Science articles every day. Follow to join our 3.5M+ monthly readers.

Following

Written by Adham Khaled

1.1K followers · 96 following

Embedded Systems Engineer || AI & Tech enthusiast || https://linktr.ee/adhamhidawy

No responses yet

Bgerby

What are your thoughts?

More from Adham Khaled and Artificial Intelligence in Plain English

In AI Mind by Adham Khaled

The Developer’s Co-pilot is Dead: How Factory AI’s Droid Ushers in the Age of Agent-Native...

Forget code completion—Droid autonomously handles entire software development workflows while beating Claude Code and GPT-5 on industry...

Sep 29 66

In Artificial Intelligence in Plain English by Simranjeet Singh

RAG is Hard Until I Know these 12 Techniques → RAG Pipeline to 99% Accuracy

RAG is Hard Until I Know these 12 Techniques → RAG Pipeline to 99% Accuracy. Best Blog to Scale or increase RAG Pipelines Accuracy.

Sep 27 498 8

In Artificial Intelligence in Plain English by Simranjeet Singh

OpenAI ML Engineer Interview Questions 2025

A mock interview with an OpenAI ML engineer covering LLM deployment, low-latency inference, quantization, mixed precision, and strategies.

Sep 24 140 7

In Data And Beyond by Adham Khaled

Google’s startup guide to AI agents: ADK, MCP, A2A, and Agentic RAG in practice

Oct 5 54

See all from Adham Khaled

See all from Artificial Intelligence in Plain English

Recommended from Medium

In Dare To Be Better by Max Petrusenko

Claude Skills: The $3 Automation Secret That’s Making Enterprise Teams Look Like Wizards

How a simple folder is replacing $50K consultants and saving companies literal days of work

Oct 17 376 5

Will Lockett

You Have No Idea How Screwed OpenAI Actually Is

When you find yourself in a hole, at what point do you stop digging?

4d ago 4.96K 148

In Coding Nexus by Code Coup

The Claude Skills Cookbook: Anthropic's New Context Engine Outperforms MCP??

Anthropic quietly released the Claude Skills Cookbook on GitHub. Initially, I assumed it was just another dull API document. But after...

3d ago 45

In Stackademic by Somendradev

10 Niche Developer Tools You Didn’t Know Existed

Let’s be real—the developer world moves fast. Every week, a dozen new tools launch promising to “boost productivity” or “make your life...

Oct 15 185 4

Ondřej Popelka

Generating user documentation from nothing

AI vision and reasoning to transform screen recordings into structured user guides

Oct 19 390 6

Farhad Ali

Perplexity Just Unleashed 10 FREE AI Agents That Do Your Entire Job (The “Comet” Shortcut)

I was drowning in a sea of research and repetitive tasks. This simple, free feature is like hiring a team of specialist VAs that works at...

5d ago 105 1

See more recommendations