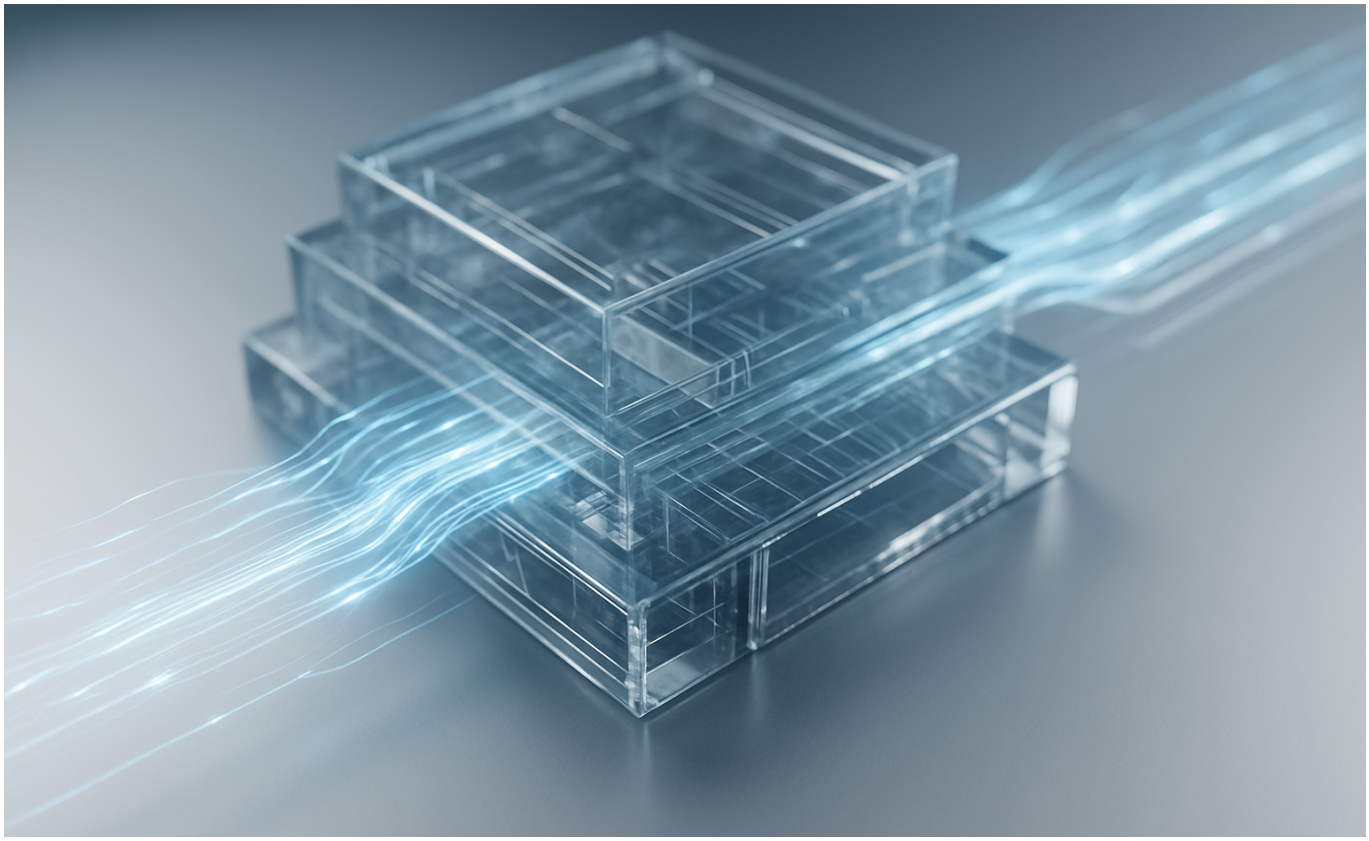
KI-4-Everyone · Daily News

17. Juli 2026



PROD

NVIDIA Vera Rubin soll KI-Training billiger machen

NVIDIAs neue Architektur Vera Rubin zielt darauf ab, den Preis pro verarbeitetem Text-Token zu senken – besonders für autonome KI-Agenten.

PROD

Bild- und Videomodelle im großen Maßstab trainieren

NVIDIA NeMo Automodel und Hugging Face Diffusers arbeiten zusammen, damit Entwickler Bild- und Videomodelle effizienter anpassen können.

PRODUKT

Nvidia bewirbt Vera Rubin als naechsten Schritt fuer agentische KI

Der Chip-Konzern stellt seine kommende Plattform als Antwort auf die teure Nachbearbeitung von KI-Modellen dar - Details bleiben im veroeffentlichten Material knapp.

Nvidia positioniert seine kommende Hardware-Generation namens Vera Rubin als Werkzeug fuer eine ganz bestimmte Phase im Leben eines KI-Modells: das sogenannte Post-Training, also die Feinschliff-Runden, nachdem ein grosses Sprachmodell bereits mit Rohdaten gefuettert wurde. Der Konzern verknuepft die Plattform in seiner Ankuendigung ausdruecklich mit dem Trend zu agentischer KI - also Systemen, die selbststaendig Aufgaben in mehreren Schritten erledigen. Die Kernbotschaft: Wer solche Agenten baut, brauche nicht nur mehr Rechenleistung, sondern vor allem guenstigere Rechenleistung pro nuetzlichem Ergebnis.

Konkret veroeffentlichte Nvidia am 17. Juli 2026 einen Blogbeitrag, in dem Vera Rubin als kommende Plattform vorgestellt wird. Als zentrale Kennzahl nennt der Konzern die Kosten pro Token - also pro erzeugter Text- oder Dateneinheit eines KI-Modells - und verspricht, diese durch ein besonders enges Zusammenspiel von Chip-Design, Software und Systemarchitektur (im Blog als 'extreme codesign' bezeichnet) zu senken. Daraus leitet Nvidia die eigentliche Verkaufsformel ab: 'Intelligence per Dollar', also Intelligenz pro eingesetztem Dollar, speziell fuer Post-Training-Workloads. Weitere technische Angaben, etwa zu Rechenleistung, Speicher oder Verfuegbarkeit, sind im vorliegenden Material nicht enthalten.

Die Einordnung ist trotzdem interessant, weil sich der Marketing-Fokus verschiebt. Lange drehte sich die Debatte um das Training riesiger Basismodelle - also die teure Grundausbildung. Mit dem Aufstieg von Agenten, die im Betrieb staendig nachjustiert, ausgerichtet und getestet werden, wird das Post-Training zum eigenen Kostenblock. Nvidia adressiert damit genau die Kundengruppe, die aktuell Milliarden in KI-Infrastruktur steckt: Cloud-Anbieter, Modellhaeuser und Unternehmen, die eigene Agenten produktiv einsetzen wollen. Wer die Kosten pro Token drueckt, senkt fuer diese Kunden die laufende Rechnung - und bindet sie zugleich staerker an die eigene Plattform. Konkurrenten, die alternative Beschleuniger bauen, geraten damit rhetorisch unter Druck, ihre eigenen Effizienzwerte offenzulegen.

Vieles bleibt allerdings offen. Der vorliegende Blogbeitrag ist eine Ankuendigung des Herstellers, keine unabhaengige Messung. Wie stark die Kosten pro Token tatsaechlich sinken, gegenueber welcher Vergleichsplattform und unter welchen Bedingungen, ist im Material nicht belegt. Auch ein konkreter Marktstart, Preise oder Kunden, die Vera Rubin bereits einsetzen, werden hier nicht genannt. Der Begriff 'Intelligence per Dollar' ist zudem eine Marketing-Metrik, keine standardisierte Groesse - vermutlich wird die Bewertung stark davon abhaengen, welche Modelle und welche Aufgaben zugrunde gelegt werden. Skepsis ist also angebracht, solange unabhaengige Benchmarks fehlen.

In den kommenden Wochen lohnt es, auf drei Dinge zu achten: erstens auf konkrete technische Daten zu Vera Rubin, die ueber den Marketing-Rahmen hinausgehen. Zweitens auf Reaktionen von Wettbewerbern im Beschleuniger-Markt, die bisher eigene Effizienz-Argumente ins Feld fuehren. Und drittens auf Aussagen grosser Modellhaeuser oder Cloud-Anbieter, ob sie Vera Rubin fuer Post-Training tatsaechlich einplanen - denn erst deren Zusagen wuerden aus einer Ankuendigung ein belastbares Signal machen.

REG

Xi grenzt sich ab: Chinas KI-Stack soll USA mit Huawei-Hardware überholen

Staatspräsident Xi positioniert Chinas KI-Politik als eigenständigen Weg. Auf der World AI Conference zeigt sich, dass der chinesische KI-Stack auf Huawei-Hardware setzt. Ein Überholmanöver gegenüber den USA wird damit für möglich gehalten.

PROD

Moonshot AI bringt Kimi K3: Chinas neues Sprachmodell fordert OpenAI heraus

Das chinesische Start-up Moonshot AI erhöht den Druck auf US-Konkurrenten. Kimi K3 ist Chinas Antwort auf Modelle von OpenAI und Anthropic. Wie Washington und das Silicon Valley reagieren, bleibt unklar.

MARKT

OpenAI-CFO stellt KI-Scorecard vor: ROI messbar machen

OpenAIs CFO Sarah Friar präsentiert eine Scorecard, um den Nutzen von KI zu messen. Sie misst Kosten pro erfolgreich erledigter Aufgabe, Zuverlässigkeit und Return on Compute. Damit soll KI-Investitionen ein konkreter Maßstab gegeben werden.

OS

LM Studio Bionic: KI-Agent für lokale Open-Source-Modelle

LM Studio bringt mit Bionic einen KI-Agenten speziell für offene Modelle. Nutzer können damit lokal laufende Modelle als Agenten einsetzen. Details zu Funktionsumfang und unterstützten Modellen sind im Material nicht genannt.

OS

Mozilla analysiert den Stand von Open-Source-KI

Mozilla veröffentlicht eine Bestandsaufnahme zur Lage offener KI-Modelle. Der Bericht beleuchtet, wo Open-Source-KI heute steht. Weitere Inhalte sind im Material nicht aufgeführt.

PROD

Musikvideo für 100 Dollar: Claude Fable 5 gegen GPT-5.6 Sol im Vergleich

Ein Musikvideo entstand mit einem Budget von 100 Dollar und zwei KI-Systemen. Claude Fable 5 und GPT-5.6 Sol traten dabei gegeneinander an. Welches System besser abschnitt, nennt das Material nicht.

PROD

Capital One setzt KI-Agenten zur Sicherheitsprüfung von Code ein

Capital One hat VulnHunter entwickelt, ein agentisches KI-Tool für Code-Sicherheit. Das System sucht selbstständig nach Schwachstellen im Quellcode. Weitere technische Details sind im Material nicht genannt.

SAFE

KI trifft Kryptografie: Schwachstellen in OpenVMs ZkVM entdeckt

KI-Systeme wurden eingesetzt, um Schwachstellen in OpenVMs Zero-Knowledge-VM zu finden. Das Projekt untersucht, was KI in formaler Verifikation und Kryptografie leisten kann. Genaue Ergebnisse sind im Material nicht spezifiziert.

TOOLS & HANDS-ON

OS

Google veröffentlicht DiffusionGemma: Bilder und Text kombiniert

DiffusionGemma-26B-A4B-it nimmt Bilder und Text als Eingabe und antwortet in Textform. Mit 1,8 Millionen Downloads ist es bereits stark gefragt.

OS

Mistral aktualisiert sein kleines Modell: Ministral-3B in neuer Version

Ministral-3-8B-Instruct-2512 ist eine aktualisierte Fassung von Mistrals kompaktem Sprachmodell – gedacht für schnelle, ressourcenschonende Anwendungen.

RES

Microsoft bringt KI-Modell für Chemie und Quantenberechnungen

Skala-1.1 richtet sich an Chemie-Forschung und unterstützt Dichtefunktionaltheorie – ein Verfahren zur Berechnung von Molekülstrukturen. Für den Alltag kaum relevant.

RES

Microsoft CoIIPRI erkennt Bildinhalte ohne vorheriges Training auf Kategorien

Das Modell klassifiziert Bilder in Kategorien, die es nie explizit gelernt hat – sogenannte Zero-Shot-Klassifikation. Das kann nützlich sein, wenn Bilddaten fehlen.

OS

Google TAPNet verfolgt beliebige Punkte in Videos automatisch

TAPNet trackt einzelne Punkte durch Videosequenzen – etwa um Bewegungen zu analysieren. Das Modell ist offen verfügbar, wurde bisher aber noch nicht heruntergeladen.

PROD

Timeline Scan: App erkennt und korrigiert Daten auf eingescannten Fotos

Das Tool liest eingescannte Fotos und versucht, Aufnahmedaten automatisch zu erkennen und zu ergänzen. Praktisch für alle, die alte Fotoalben digitalisieren.

AUSBLICK

Keine Termine gemeldet.