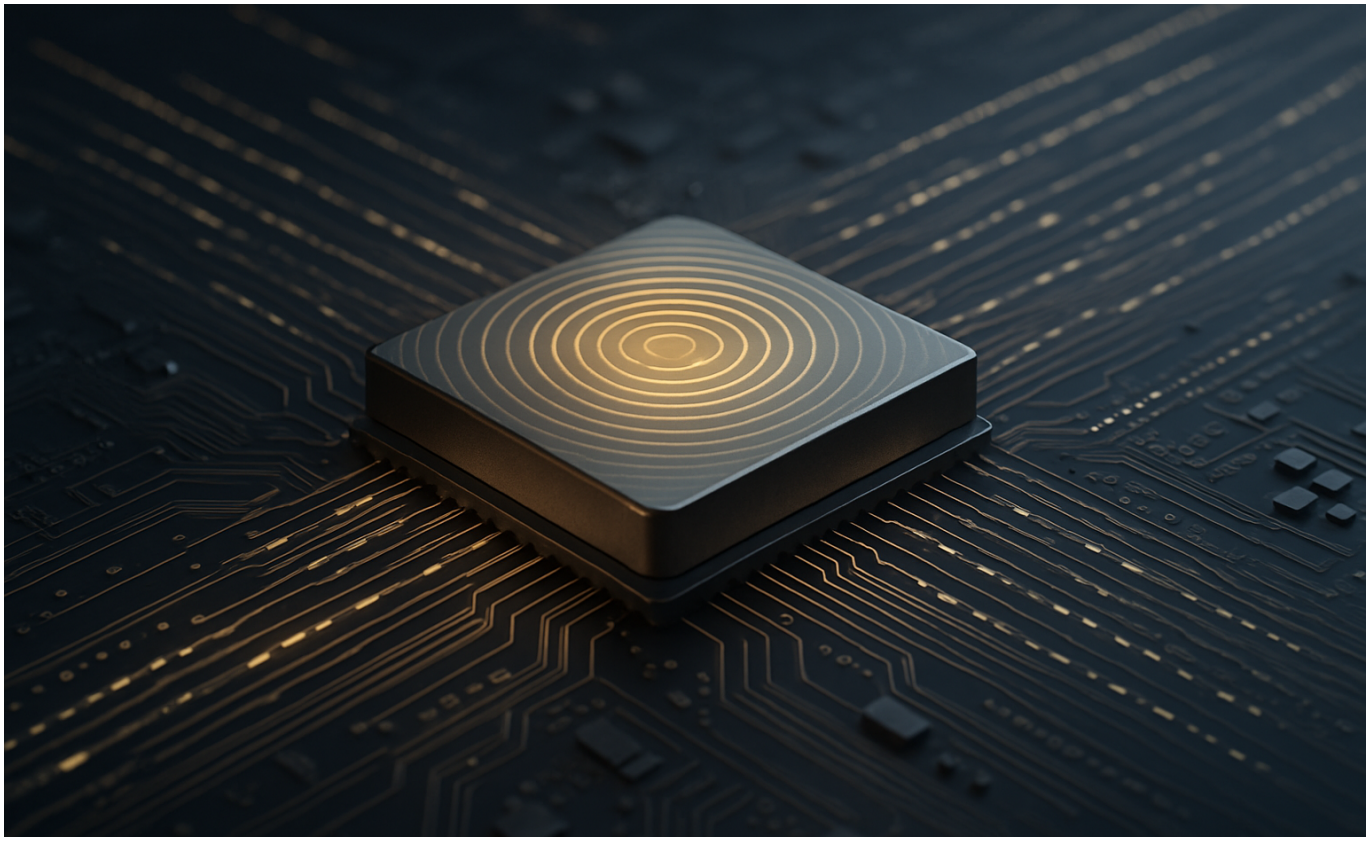
KI-4-Everyone · Daily News

24. August 2026



PROD

NVIDIA Vera Rubin NVL72 liefert 30x mehr Leistung pro Watt

KI-Agenten brauchen laut OpenRouter-Daten 15x mehr Rechenaufwand als ein normaler Chat. NVIDIAs neuer Chip soll das deutlich effizienter machen.

PROD

NVIDIA erweitert Vera Rubin für schnellere KI-Agenten

Groq 3 LPX ist jetzt in der Vollproduktion. NVIDIA kombiniert es mit dem Vera Rubin NVL72, um KI-Agenten schneller mit Antworten zu versorgen.

PRODUKT

Nvidias neuer Server-Riese Vera Rubin NVL72: Bis zu 30-mal mehr Arbeit pro Watt fuer KI-Agenten

Nvidia positioniert seine kommende Rechenplattform als Antwort auf den wachsenden Stromhunger autonomer KI-Systeme, die deutlich mehr Rechenzeit verschlingen als klassische Chatbots.

Ein einzelner Klick auf einen Chatbot ist eine Sache. Ein KI-Agent, der im Hintergrund eine Firma durchleuchtet, Datenbanken abfragt, Nachrichten und Geschaeftsberichte durchsucht und am Ende eine Investitionsempfehlung zusammenstellt, ist eine ganz andere. Genau an dieser Verschiebung setzt Nvidia mit seiner neuen Plattform Vera Rubin NVL72 an - und verspricht, die Effizienz solcher Agenten-Arbeitslasten drastisch zu steigern.

In einem Blogbeitrag vom 24. August 2026 kuendigt Nvidia die Plattform Vera Rubin NVL72 an und stellt sie als neuen Effizienzmassstab fuer sogenannte agentische KI dar - also KI-Systeme, die eigenstaendig Teilaufgaben planen, Werkzeuge nutzen und Unter-Agenten (kleinere KI-Helfer fuer Teilaufgaben) aufrufen. Laut Nvidia soll das System bis zu 30-mal mehr Arbeit pro Watt leisten. Als Beleg fuer den wachsenden Bedarf zitiert Nvidia Daten von OpenRouter, einem Vermittlungsdienst fuer KI-Modelle, wonach agentische Arbeitslasten rund 15-mal so viele Tokens verbrauchen wie eine einfache Chat-Anfrage. Ein Token ist dabei eine kleine Text-Einheit, mit der KI-Modelle rechnen.

Der Zeitpunkt ist kein Zufall. Waehrend klassische Chatbots seit Jahren im Einsatz sind, verschiebt sich die Branche gerade in Richtung autonomer Agenten, die deutlich mehr Rechenschritte pro Aufgabe brauchen. Damit steigen Stromverbrauch und Kuehlbedarf in Rechenzentren spuerbar - ein Thema, das Betreiber, Energieversorger und zunehmend auch die Politik unter Druck setzt. Nvidia, das den Markt fuer KI-Beschleuniger dominiert, versucht mit dem Effizienzargument zwei Fliegen zu schlagen: Es adressiert die Kostenrechnung der Cloud-Kunden und zugleich die oeffentliche Debatte um den Energiehunger von KI. Wer eine Plattform anbieten kann, die pro Watt mehr Ergebnisse liefert, hat im Werben um Hyperscaler und Unternehmenskunden ein starkes Verkaufsargument.

Offen bleibt allerdings vieles. Der Wert von 'bis zu 30x mehr Arbeit pro Watt' ist eine Herstellerangabe - unklar ist, gegen welche Vorgaengergeneration genau verglichen wird, unter welchen Lastprofilen und mit welchen Modellen. Auch ob sich die Effizienzgewinne in realen Agenten-Anwendungen so einstellen wie in Nvidias Benchmarks, laesst sich aus dem Material nicht ableiten. Die OpenRouter-Zahl zum 15-fachen Token-Verbrauch stuetzt zwar die Erzaehlung vom wachsenden Bedarf, sagt aber nichts darueber aus, wie repraesentativ die dort gemessenen Workloads fuer den Gesamtmarkt sind. Und schliesslich: Ein Termin fuer die breite Verfuegbarkeit von Vera Rubin NVL72 ist im vorliegenden Beitrag nicht belegt.

In den kommenden Wochen lohnt es sich zu beobachten, ob unabhaengige Tester oder grosse Cloud-Anbieter die Effizienzangaben nachvollziehen - und ob Konkurrenten wie AMD oder spezialisierte KI-Chip-Startups mit eigenen Zahlen kontern. Denn die eigentliche Frage hinter der Ankuendigung lautet nicht, ob KI-Agenten kommen, sondern wer die Rechenzentren baut, in denen sie bezahlbar laufen.

MARKT

Alibaba bringt KI-Videogenerator und kündigt Milliarden-Investitionen an

Alibaba hat ein neues KI-Videomodell veröffentlicht und gleichzeitig massive Infrastrukturausgaben angekündigt. Damit stärkt der Konzern seine Position im globalen KI-Wettbewerb. Konkrete Investitionssummen nennt das Material nicht.

RES

Warnung: KI-Abhängigkeit könnte Programmier-Kompetenz aushöhlen

Wer dauerhaft auf KI-Tools setzt, verlernt möglicherweise das eigenständige Coden. Diese These steht im Mittelpunkt des Beitrags. Ob und wann das messbar eintritt, lässt das Material offen.

OS

Entwickler baut KI-Begleiter, der Skyrim in Echtzeit mitspielt

Ein Entwickler hat einen latenzarmen KI-Companion gebaut, der das Spiel Skyrim gemeinsam mit ihm spielt. Das System reagiert in Echtzeit auf das Spielgeschehen. Technische Details zur Umsetzung nennt das Material nicht.

REG

Googles A2A und Anthropics MCP vereint unter Linux-Foundation-Dach

Googles A2A-Protokoll tritt der Agentic AI Foundation der Linux Foundation bei, die über 250 Mitglieder zählt. Damit kommen A2A und Anthropics MCP unter eine neutrale Open-Governance. Das ist ein Schritt zur Standardisierung von KI-Agenten-Schnittstellen.

RES

FDA lässt Bluttest zur Alzheimer-Früherkennung zu

Die FDA hat einen Bluttest freigegeben, der bei der Bewertung von Alzheimer helfen soll. Ob KI bei der Auswertung eine Rolle spielt, geht aus dem Material nicht hervor. Der Test ergänzt bestehende Diagnoseverfahren.

PROD

Nvidia: KI-Fabriken brauchen Gesamtarchitektur, keine Einzelbeschleuniger

Nvidia argumentiert, dass moderne KI-Infrastruktur als vollständige Fabrik gebaut sein muss. Entscheidende Kennzahlen sind Token pro Sekunde, Token pro Watt und Kosten pro Token. Einzelne Beschleuniger reichen demnach nicht aus.

MARKT

HN-Diskussion: Warum treffen Unternehmensfehler oft die Falschen?

Ein Nutzer schildert, wie seine Partnerin nach 15 Jahren und über 30 Mitarbeitenden in der Führung entlassen wurde. Die Diskussion fragt, warum Entlassungswellen häufig leistungsstarke Mitarbeitende treffen. Einen KI-Bezug enthält das Material nicht explizit.

PROD

"Wir nutzen KI für gar nichts" – ein Unternehmen erklärt warum

Ein Team berichtet, dass es KI-Tools konsequent aus seinen Prozessen heraushält. Gründe und Kontext nennt das Material nicht im Detail. Die Aussage steht im Gegensatz zum allgemeinen Branchentrend.

TOOLS & HANDS-ON

PROD

OpenAI bringt Verlaufs-Gedächtnis für Mac-Nutzer – ähnlich wie Windows Recall

Business-, Enterprise- und Pro-Nutzer können ihre gesamte Rechnerarbeit auf macOS von ChatGPT mitschreiben lassen. Die gespeicherten Memory-Files liegen offen zugänglich ab.

PROD

The Blood of Dawnwalker: Entwickler kündigt 60-fps-Modus nach Kritik an

Rebel Wolves hatte für Konsolen zunächst maximal 40 fps angekündigt. Nach Protesten aus der Community gibt es nun doch einen 60-fps-Modus.

PROD

FRITZ!Smart Thermo 302 bei Amazon für 49 Euro erhältlich

Der Heizkörperregler regelt die Raumtemperatur automatisch und lässt sich per App steuern. Amazon bietet ihn aktuell reduziert an.

PROD

Lego Saurons Helm bei Amazon auf 45,99 Euro gesenkt

Laut Preistracker war das Herr-der-Ringe-Set beim Onlinehändler noch nie günstiger. Amazon listet es aktuell zum Tiefstpreis.

PROD

Microsoft-App setzt Bing als Standardsuche in allen erkannten Browsern

Die App erkennt installierte Browser und wechselt dort die Standardsuchmaschine zu Bing. Außerdem installiert sie passende Browsererweiterungen automatisch.

OS

Ansible-Workshop: Konfigurationsverwaltung ohne Agenten in der Praxis

Der Workshop vermittelt Playbooks und Rollen für den Praxiseinsatz mit Ansible. Das Tool gilt als agentenlose Alternative zu Puppet, Chef und Saltstack.

AUSBLICK

Keine Termine gemeldet.