KI-4-Everyone · Daily News

5. Mai 2026

RES

Microsoft zeigt neue Technik für riesige vernetzte Rechenzentren

Microsoft präsentiert auf der NSDI '26 Forschungsergebnisse zu verteilten Systemen, Netzwerken und deren Verbindung mit KI.

PROD

NVIDIA und ServiceNow bringen KI-Agenten in Unternehmen

KI soll nicht mehr nur antworten, sondern selbstständig handeln. NVIDIA und ServiceNow arbeiten gemeinsam an Agenten, die komplexe Aufgaben im Unternehmensalltag übernehmen.

FORSCHUNG

Microsoft auf der NSDI 2026: Wenn KI das Rechenzentrum umbaut

Microsoft Research zeigt auf der Netzwerk-Konferenz NSDI, wie sich grosse verteilte Systeme veraendern - getrieben auch durch die Anforderungen von KI-Workloads.

Wer ueber KI redet, redet meistens ueber Modelle, Chatbots oder Bilder. Doch hinter jedem Sprachmodell stehen riesige Rechenzentren, in denen tausende Computer gleichzeitig zusammenarbeiten muessen. Microsoft Research nutzt jetzt eine Fachkonferenz, um zu zeigen, wie sehr KI gerade die Technik unter der Motorhaube veraendert - also genau dort, wo die meisten Nutzer nie hinschauen.

Anlass ist die NSDI 2026, eine etablierte Konferenz fuer vernetzte Systeme (NSDI steht fuer Networked Systems Design and Implementation, ein Treffen von Forschern, die sich mit der Architektur grosser Computernetze beschaeftigen). In einem Blogbeitrag vom 5. Mai 2026 kuendigt Microsoft Research an, dort eigene Arbeiten zu praesentieren. Thematisch geht es laut Ankuendigung um Fortschritte beim Bau und Betrieb von grossen verteilten Systemen - das umfasst Rechenzentren, Netzwerktechnik und die wachsende Schnittstelle zur kuenstlichen Intelligenz. Welche konkreten Papiere Microsoft einbringt, welche Autoren beteiligt sind und welche Ergebnisse im Detail vorgestellt werden, geht aus dem hier vorliegenden Material nicht hervor. Klar ist nur die Stossrichtung: Netzwerke, Datacenter, KI - und ihre Verbindung untereinander.

Warum ist das relevant, auch fuer Leser, die nie ein Rechenzentrum von innen sehen? Weil die juengste KI-Welle die Anforderungen an die Infrastruktur sprunghaft erhoeht hat. Sprachmodelle trainieren ueber tausende Grafikprozessoren hinweg, die im Bruchteil einer Sekunde miteinander reden muessen. Wenn das Netzwerk dazwischen klemmt, steht teure Hardware still. Dass ein Konzern wie Microsoft - selbst grosser Cloud-Anbieter und eng mit OpenAI verflochten - seine Forschung in diesem Bereich oeffentlich macht, ist daher mehr als akademisches Schaulaufen. Es ist auch ein Signal an Kunden und Konkurrenten, dass die naechste Wettbewerbsrunde im KI-Geschaeft nicht nur ueber Modelle entschieden wird, sondern ueber das, was sie traegt: Leitungen, Switches, Protokolle, Betriebsverfahren. Wer hier effizienter arbeitet, kann KI-Dienste billiger oder schneller anbieten.

Offen bleibt im vorliegenden Material vieles. Es ist nicht belegt, welche konkreten Probleme Microsoft auf der NSDI loest - etwa ob es um Ueberlastsituationen beim Training, um Ausfallsicherheit, um Energieeffizienz oder um neue Hardware-Software-Kombinationen geht. Auch Zahlen zu Leistungsgewinnen, getesteten Systemgroessen oder Vergleiche mit anderen Anbietern nennt der Blogbeitrag in der hier vorliegenden Form nicht. Vermutlich werden die Details erst mit den eigentlichen Konferenzbeitraegen sichtbar. Auch das Datum wirkt erklaerungsbeduerftig: Die Veroeffentlichung traegt das Datum 5. Mai 2026, die Einordnung dazu ist im Material nicht weiter belegt.

Wer das Thema verfolgen will, sollte auf zwei Dinge achten: erstens die einzelnen Forschungspapiere, die Microsoft im Umfeld der NSDI veroeffentlichen duerfte und die konkretere Aussagen erlauben werden; zweitens darauf, ob andere grosse Cloud-Anbieter aehnliche Beitraege liefern. Denn die spannende Frage ist nicht, ob KI die Rechenzentren veraendert - das tut sie bereits -, sondern wer am schnellsten lernt, diese Veraenderung in laufenden Betrieb zu uebersetzen. Im hier verfuegbaren Material ist das noch nicht zu beantworten.

RES

IBM Granite 4.1: Kleines Modell schlägt deutlich größere Konkurrenten

IBM hat Granite 4.1 mit 8 Milliarden Parametern veröffentlicht. Laut Benchmarks erreicht es die Qualität von 32B-MoE-Modellen. Für Unternehmen ist das relevant, weil kleinere Modelle günstiger zu betreiben sind.

REG

Google Chrome installiert 4-GB-KI-Modell ohne Nutzereinwilligung

Chrome lädt laut einem Bericht still ein 4 GB großes KI-Modell auf dein Gerät – ohne Zustimmung zu fragen. Das wirft Fragen zur Datenkontrolle und Transparenz auf.

PROD

Anthropic zeigt, wie KI-Agenten im Finanzsektor eingesetzt werden

Anthropic beschreibt konkrete Anwendungsfälle für KI-Agenten in Finanzdienstleistungen. Welche Aufgaben die Agenten übernehmen sollen, geht aus dem Titel hervor – Details sind im Material nicht enthalten.

MARKT

AI Product Graveyard: Eine Sammlung gescheiterter KI-Produkte

Ein Projekt dokumentiert KI-Produkte, die bereits eingestellt wurden. Wie viele Einträge die Liste umfasst oder wer dahintersteckt, ist im Material nicht enthalten.

MARKT

Wenn KI-Tools da sind, aber das Unternehmen nichts dazulernt

Ein Beitrag beschreibt das Problem: Viele Mitarbeitende nutzen KI, doch das Wissen bleibt individuell und verbreitet sich nicht im Unternehmen. Organisationales Lernen findet so nicht statt.

PROD

Datenbankfehler durch KI? Meist steckt ein Mensch dahinter

Ein Beitrag argumentiert, dass KI nicht für gelöschte Datenbanken verantwortlich ist – sondern der Mensch, der den Befehl erteilt hat. Die Frage der Verantwortung bleibt also beim Nutzer.

RES

Drei inverse Gesetze der KI: Was wächst, wenn Modelle besser werden?

Ein Beitrag formuliert drei gegenläufige Gesetzmäßigkeiten rund um KI-Entwicklung. Welche Thesen genau aufgestellt werden, ist im vorliegenden Material nicht enthalten.

TOOLS & HANDS-ON

PROD

GPT-5.5 Instant: OpenAIs neues Standard-Modell in ChatGPT

OpenAI ersetzt das bisherige Standardmodell in ChatGPT durch GPT-5.5 Instant. Es soll genauere Antworten liefern, weniger halluzinieren und sich besser an persönliche Einstellungen anpassen.

PROD

ChatGPT bekommt einen Anzeigenmanager für Werbetreibende

OpenAI startet eine Beta-Version eines selbstverwalteten Ads Managers für ChatGPT – mit Klickpreis-Geboten und Messtools. Gespräche und Werbung sollen laut OpenAI getrennt bleiben.

PROD

GPT-5.5 führt Benchmarks an – halluziniert aber weiterhin

Laut The Batch übertrifft GPT-5.5 andere Modelle in Leistungstests, produziert aber noch immer falsche Aussagen. Daneben führt Kimi K2.6 die Liste der offenen Sprachmodelle an.

OS

Mistral Medium 3.5 mit 128 Milliarden Parametern ist offen verfügbar

Mistral veröffentlicht ein großes Sprachmodell mit 128 Milliarden Parametern – es läuft unter anderem mit vLLM und richtet sich an englischsprachige Anwendungen. Bereits über 15.000 Downloads auf Hugging Face.

OS

Googles Gemma 4 (31B) versteht Text und Bilder gleichzeitig

Google stellt eine Variante seines Gemma-4-Modells bereit, die verschiedene Eingabetypen verarbeiten kann – also nicht nur Text, sondern auch andere Datenformate. Das Modell ist offen auf Hugging Face verfügbar.

OS

Airbyte Agents: KI-Agenten greifen auf viele Datenquellen zu

Airbyte startet ein Tool, das KI-Agenten Kontext aus mehreren Datenquellen gleichzeitig liefert – aufgebaut auf sechs Jahren Erfahrung mit Daten-Konnektoren. Das Projekt ist Open Source.

AUSBLICK

Keine Termine gemeldet.