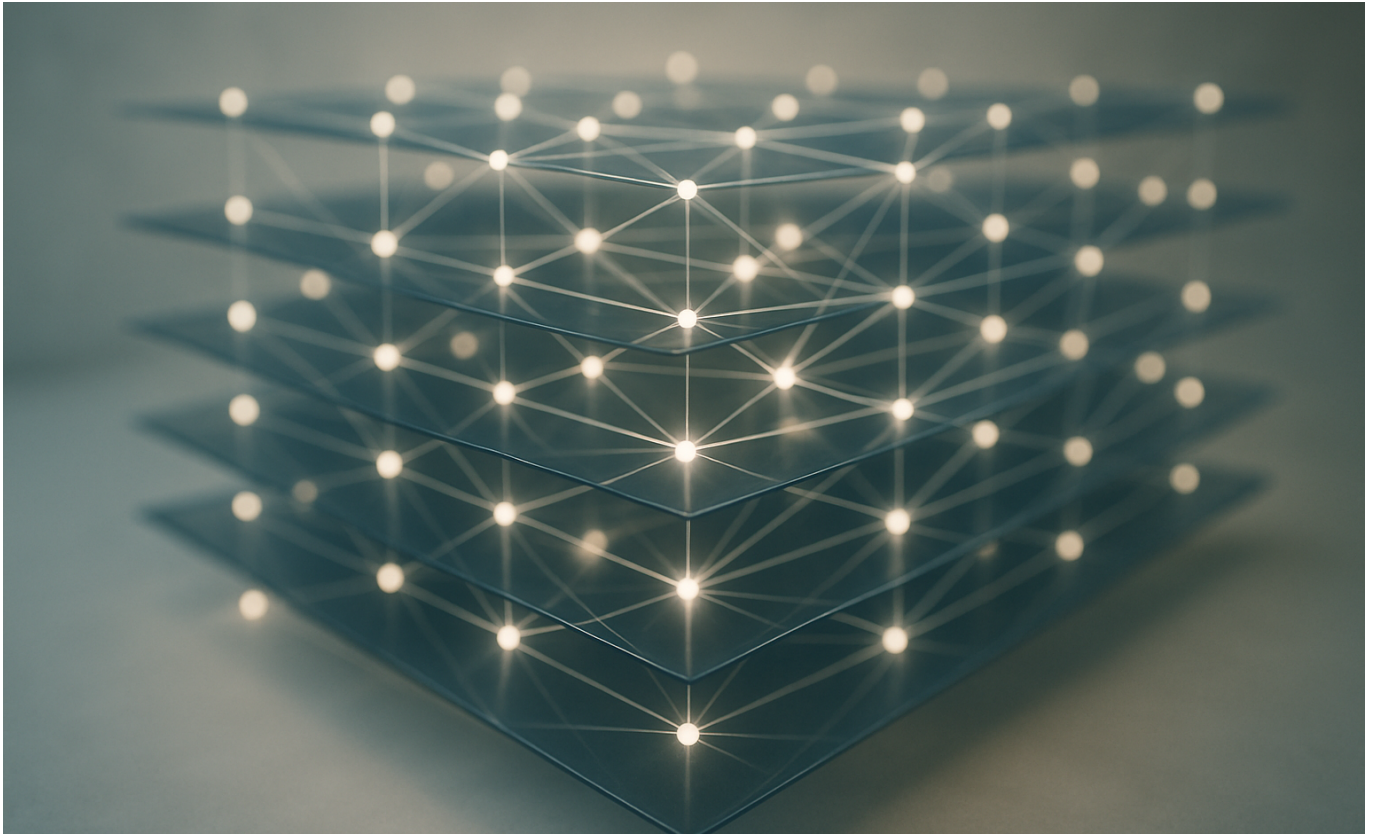


KI-4-Everyone · Daily News

15. September 2026



PROD

Google stellt Gemini 3.8 Live mit erweitertem Denken vor

Google bringt zwei neue Modelle: Gemini 3.8 Live und Gemini 3.8 Live Extended Thinking. Letzteres soll komplexere Aufgaben durch längeres Nachdenken lösen.

PROD

NVIDIA zeigt, wie KI-Rechenzentren weniger Strom verbrauchen

NVIDIA-Manager Ian Buck erklärte auf dem AI Infra Summit, wie optimierte Token-Verarbeitung pro Watt KI-Fabriken effizienter macht.

Google kuendigt Gemini 3.8 Live an - mit einem zweiten, langsamer denkenden Modus

Der Konzern stellt zwei Varianten seines Live-Modells vor. Zu Details, Preisen oder Verfuegbarkeit liefert das Material bislang kaum belastbare Informationen.

Google hat mit Gemini 3.8 Live und einer zweiten Variante namens 3.8 Live Extended Thinking zwei neue Ausbaustufen seines Sprachmodells angekuendigt. Bemerkenswert ist weniger die schiere Existenz einer weiteren Versionsnummer, sondern die Aufspaltung selbst: Ein Modell soll offenbar schnell antworten, das andere sich mehr Zeit fuer laengere Ueberlegungen nehmen. Damit greift Google ein Muster auf, das andere Anbieter zuletzt ebenfalls verfolgt haben - naemlich Nutzern die Wahl zu lassen zwischen Tempo und Gruendlichkeit.

Konkret nennen die vorliegenden Quellen zwei Namen: Gemini 3.8 Live und Gemini 3.8 Live Extended Thinking. Die Ankuendigung stammt laut Material vom 15. September 2026, veroeffentlicht ueber den DeepMind-Blog und aufgegriffen im KI-Bereich von Hacker News. Weitere Fakten - etwa welche Faehigkeiten die beiden Modelle konkret unterscheiden, ob es sich um multimodale Systeme handelt, welche Kontextfenster sie bieten oder ab wann Entwickler und Endkunden Zugriff bekommen - gehen aus den beiden Titeln nicht hervor. Der Begriff Live legt nahe, dass es um Echtzeitinteraktion geht, etwa in Sprache oder Video, doch belegen laesst sich das aus dem Material nicht.

Die Einordnung faellt trotzdem nicht schwer: Google befindet sich in einem oeffentlich sichtbaren Wettlauf mit OpenAI, Anthropic und einer wachsenden Zahl chinesischer Anbieter. Der Zusatz Extended Thinking spiegelt einen Trend wider, den man seit einiger Zeit bei mehreren Anbietern beobachtet - Modelle, die intern laenger rechnen, bevor sie antworten, gelten als staerker bei komplexen Aufgaben wie Mathematik oder Programmierung. Dass

Google beide Ansaetze parallel unter derselben Versionsnummer bundelt, koennte darauf hindeuten, dass der Konzern seinen Nutzern die Entscheidung ueber den Trade-off zwischen Geschwindigkeit und Antwortqualitaet aktiv in die Hand geben will. Fuer Entwickler, die Anwendungen auf Basis von Gemini bauen, waere das eine wichtige Weichenstellung: Ein Chat-Assistent braucht andere Antwortzeiten als ein Analysewerkzeug.

Vieles bleibt allerdings offen. Das Material enthaelt keine Angaben zu Benchmarks, zu Preismodellen, zu Sprachunterstuetzung oder zu regionaler Verfuegbarkeit - insbesondere nicht fuer Europa, wo regulatorische Fragen den Start neuer KI-Produkte regelmaessig verzoegern. Auch ob 3.8 eine echte Zwischenversion zu einem kommenden Gemini 4 ist oder eher ein iteratives Update, laesst sich aus den vorliegenden Titeln nicht ableiten. Ob die Extended-Thinking-Variante teurer sein wird oder langsamer antwortet, ist ebenfalls nicht im Material belegt. Wer belastbare Details braucht, sollte den offiziellen DeepMind-Blogbeitrag abwarten beziehungsweise dessen Inhalt pruefen, sobald er zugaenglich ist.

In den kommenden Tagen lohnt der Blick auf drei Punkte: Erstens, ob unabhengige Tester Benchmarks veroeffentlichen, die den Unterschied zwischen den beiden Modi quantifizieren. Zweitens, ob und wie schnell die neuen Modelle in Google-Produkten wie der Gemini-App oder in der Google-Cloud-Plattform Vertex AI auftauchen. Und drittens, wie Wettbewerber reagieren - denn eine geteilte Modellstrategie mit schneller und langsamer Variante koennte, wenn sie sich bewaehrt, zum neuen Standard in der Branche werden.

REG

Sanders und Bannon fordern gemeinsam KI-Bremsen in Washington

Eine politisch ungewöhnliche Koalition trat bei der 'Pro-Human Assembly' in Washington auf. Bernie Sanders und Steve Bannon verlangten gemeinsam neue KI-Regulierung. Das Signal könnte die US-KI-Policy-Debatte merklich verschieben.

PROD

Neue KI-Modellreihe: System One Models und Jev vorgestellt

Ein Anbieter hat die Modellreihe 'System One' sowie das Modell 'Jev' eingeführt. Details zu Leistung und Einsatzbereich sind im Material nicht weiter ausgeführt. Weitere Informationen zu Preisen oder Verfügbarkeit: unklar.

SAFE

KI-Agenten schaden dem Internet - Kritik wird lauter

Autonom handelnde KI-Agenten schaden dem offenen Internet zunehmend. Der Beitrag argumentiert, dass diese Systeme Strukturen kaputt machen, auf denen das Web aufgebaut ist. Konkrete Gegenmaßnahmen nennt das Material nicht.

REG

Ex-FTC-Chefin Khan: KI-CEOs sollten verhaftet werden können

Lina Khan, frühere FTC-Chefin, fordert härtere Strafen für KI-Manager und zieht dafür ein Gesetz aus dem Jahr 1934 heran. Sie plädiert dafür, Handschellen als Mittel der Regulierung ernstzunehmen. Die genauen rechtlichen Mechanismen: unklar.

PROD

NVIDIA zeigt, wie KI-Fabriken Stromverbrauch flexibel steuern

Silicon Valley Power schickte an einem Augustabend ein Signal an eine KI-Fabrik, den Stromverbrauch anzupassen. Varun Sivaram beobachtete das per Zoom mit rund 40 weiteren Personen vom Team Emerald AI aus San Francisco. NVIDIA beschreibt, wie solche 'AI Factories' Megawatt-Lasten in Token-Output umrechnen.

RES

Google Research: Neue Methode beschleunigt KI-Suche bei komplexen Anfragen

Google Research stellt den Ansatz 'Retrieve-for-Train' vor, der Engpässe bei der KI-Inferenz umgehen soll. Die Methode zielt auf komplexe Suchanfragen, bei denen bestehende Systeme langsam werden. Konkrete Benchmarkzahlen nennt das Material nicht.

RES

KI-Agent besteht eine Aufgabe - aber schafft er sie nochmal?

Forschende stellen die Frage, ob KI-Agenten Aufgaben zuverlässig wiederholen können oder nur zufällig erfolgreich sind. Einmaliger Erfolg sagt wenig über echte Kompetenz aus. Das ist ein grundlegendes Problem beim Testen von KI-Agenten.

RES

KI unterläuft unsere bisherigen Maßstäbe für Expertise

KI-Systeme knacken Tests und Kennzahlen, die früher als Beweis für menschliche Kompetenz galten. Das macht es schwerer zu beurteilen, wann ein System wirklich etwas kann. Der Beitrag diskutiert, welche Proxys für Expertise noch taugen.

PROD

Cartesian: KI-Tool entwirft 3D-Modelle direkt aus Designvorgaben

Cartesian richtet sich an Designer, die ohne manuelle Modellierung 3D-Objekte erstellen wollen. Das Tool nutzt KI, um aus Eingaben fertige 3D-Geometrien zu erzeugen.

PROD

Kinderklinik setzt Open-Source-KI von NVIDIA für Herzdiagnosen ein

Ein großes Kinderkrankenhaus nutzt NVIDIA-KI-Tools zur Analyse von Herzerkrankungen. Die eingesetzte Software basiert auf Open-Source-Komponenten.

PROD

MediaTek Dimensity 9600 Pro: Neuer Smartphone-Chip im 2-nm-Verfahren

MediaTek stellt den Dimensity 9600 Pro vor – gefertigt von TSMC im 2-nm-Verfahren mit Arms neuen C2-Kernen. Der Chip soll deutlich mehr Leistung in Smartphones bringen.

PROD

Meta One: Abo-Pakete bündeln Social-Media-Apps mit KI-Nutzung

Meta bietet ab sofort Abo-Bundles an, die App-Abonnements mit erweiterter KI-Nutzung kombinieren. Die Pakete sind weltweit in mehreren Preisstufen verfügbar.

OS

Sunk Cost: Rechner zeigt, wann sich lokale KI-Hardware amortisiert

Das Tool vergleicht die Hardwarekosten eines eigenen KI-Rechners mit den laufenden Kosten für Cloud-Modelle. Du gibst Modell, Gerät und tägliche Token-Nutzung ein – und bekommst die Amortisationsdauer.

PROD

Agilys neuer Humanoidroboter weicht Menschen aktiv aus

Der Roboter erkennt menschliche Kollegen und hält von sich aus an oder geht in die Hocke, um Verletzungen zu vermeiden. Damit kann er erstmals ohne Schutzkäfig oder Sicherheitsbarrieren eingesetzt werden.

Keine Termine gemeldet.