KI-4-Everyone · Daily News

10. Juni 2026

OS

DiffusionGemma macht lokale KI bis zu 4-mal schneller

Google DeepMind hat DiffusionGemma veröffentlicht – ein offenes Modell für besonders schnelle Textgenerierung. NVIDIA hat es zusätzlich für GeForce-RTX-Grafikkarten optimiert.

SAFE

Robotaxis: Sicherheit muss von Anfang an eingebaut sein

Robotaxi-Dienste fahren bereits in Dutzenden Städten kommerziell. Experten fordern, Sicherheit nicht nachträglich hinzuzufügen, sondern von Grund auf einzuplanen.

OPEN SOURCE

DiffusionGemma: Google DeepMind bringt KI-Text in Blöcken statt Wort für Wort

Ein neues Open-Source-Modell soll lokal vier Mal schneller laufen als bisherige Sprachmodelle - indem es eine Technik aus der Bildgenerierung auf Text überträgt.

Sprachmodelle tippen heute wie ein Mensch, der jedes Wort einzeln in die Tastatur hackt: erst eins, dann das nächste, dann das übernächste. Google DeepMind stellt nun ein Modell vor, das anders arbeitet - es schiebt ganze Textblöcke gleichzeitig heraus. Das Ergebnis heisst DiffusionGemma, ist offen verfügbar und soll auf dem heimischen Rechner rund vier Mal schneller antworten als vergleichbare Modelle. Damit verschiebt sich, was 'lokale KI' überhaupt bedeuten kann.

Veröffentlicht wurde DiffusionGemma laut den vorliegenden Berichten am 10. Juni durch Google DeepMind als experimentelles offenes Modell. Es nutzt einen Diffusionsansatz - eine Technik, die bislang vor allem aus der Bildgenerierung bekannt ist, wo ein Modell aus zufälligem Rauschen schrittweise ein fertiges Bild herausarbeitet. DeepMind überträgt dieses Prinzip auf Text: Statt Wort für Wort entsteht der Output in parallelen Blöcken. NVIDIA hat das Modell parallel für seine eigenen Chips optimiert, konkret für GeForce-RTX-Grafikkarten, die RTX-PRO-Plattform und die DGX-Spark-Systeme - also von Endkunden-PCs bis hin zu Server-Hardware.

Relevant ist das aus zwei Gründen. Erstens trifft es einen wunden Punkt heutiger KI-Nutzung: Die Wartezeit, bis ein Chatbot seine Antwort fertig herausgetippt hat, ist ein spürbarer Bremsklotz - besonders bei längeren Texten und besonders dann, wenn das Modell lokal auf einem einzelnen Rechner läuft und nicht auf einer Cloud-Farm. Zweitens passt es in einen Trend, KI weg von zentralen Rechenzentren und hin zum eigenen Gerät zu bringen. Wer lokal arbeitet, gewinnt Datenschutz und Unabhängigkeit, zahlt aber bisher mit Geschwindigkeit. Genau hier setzt DiffusionGemma an - und NVIDIA profitiert offensichtlich davon, weil schnelle lokale Modelle die eigene Hardware attraktiver machen. Konkurrenten, die auf klassische, Wort-für-Wort arbeitende Sprachmodelle setzen, geraten durch solche Ansätze unter Erklärungsdruck.

Vieles bleibt allerdings offen. Die Berichte nennen zwar einen Geschwindigkeitsvorteil um den Faktor vier, aber nicht klar im Vergleich zu welchem Modell, in welcher Aufgabe und auf welcher Hardware genau dieser Wert gemessen wurde. Auch zur Textqualität - also ob die parallel erzeugten Blöcke inhaltlich genauso konsistent sind wie klassisch generierter Text - liefert das Material keine Belege. Diffusionsansätze für Sprache gelten generell als noch nicht so ausgereift wie für Bilder; ob DiffusionGemma diesen Abstand schliesst, ist im vorliegenden Material nicht belegt. Dass das Modell als 'experimentell' bezeichnet wird, deutet darauf hin, dass DeepMind selbst es noch nicht als Produktivreif einstuft.

Worauf in den nächsten Tagen zu achten ist: erste unabhängige Tests von Entwicklern, die das Modell auf eigener Hardware ausprobieren, sowie konkrete Vergleiche mit etablierten offenen Modellen wie der bisherigen Gemma-Reihe. Spannend wird auch, ob andere Anbieter nachziehen und Diffusion als Standardansatz für schnelle lokale Textmodelle etablieren - oder ob es eine Nischenlösung bleibt.

PROD

Microsoft bringt eigene KI-Modellfamilie: MAI mit 7 Modellen

Microsoft veröffentlicht erstmals eine vollständige selbst entwickelte Modellfamilie. Dazu gehören MAI-Thinking-1, MAI-Code-1-Flash und MAI-Image-2.5. Außerdem kündigt Microsoft eine Kooperation mit der Mayo Clinic für ein Healthcare-Modell an.

PROD

Anthropic kündigt Claude Fable 5 und Claude Mythos 5 an

Anthropic stellt zwei neue Modelle vor: Claude Fable 5 und Claude Mythos 5. Sie richten sich laut Anthropic auf anspruchsvolle Wissensarbeit und komplexe Coding-Aufgaben. Weitere Details zu Leistung oder Preisen sind im Material nicht enthalten.

OS

Apache Burr: Open-Source-Framework für zuverlässige KI-Agenten

Apache Burr ist ein Open-Source-Werkzeug zum Bauen von KI-Agenten und -Anwendungen. Der Fokus liegt auf Zuverlässigkeit. Weitere technische Details oder Versionsinformationen sind im Material nicht enthalten.

SAFE

Sicherheitslücke: 0,01-Euro-Überweisung soll Banking-KI kompromittieren

Eine Banküberweisung von nur 0,01 Euro kann offenbar einen KI-Agenten im Banking-Bereich manipulieren. Das zeigt, wie anfällig KI-Agenten für sogenannte Prompt-Injection-Angriffe über Eingabedaten sind. Details zum betroffenen System nennt das Material nicht.

REG

Deutsches Gericht erklärt Google für falsche KI-Antworten haftbar

Ein deutsches Gericht hat Google für fehlerhafte Antworten in seinen KI-Overviews haftbar erklärt. Das Urteil könnte weitreichende Folgen für KI-Suchdienste in Europa haben. Welches Gericht entschieden hat, nennt das Material nicht.

REG

Neue Politikdebatte: Wie soll der Staat auf KI-Beschleunigung reagieren?

Der Beitrag diskutiert politische Antworten auf das exponentielle Wachstum von KI. Konkrete Maßnahmen oder Akteure nennt das Material nicht. Das Thema zeigt, dass Regulierungsfragen rund um KI-Tempo zunehmen.

REG

Unternehmen integriert Tracker für Telefon, AirPods und Smartwatches in Kennzeichenscanner

Ein Unternehmen plant, Ortungs-Tracker für Mobiltelefone, AirPods und Smartwatches in automatische Kennzeichenlesegeräte (ALPRs) einzubauen. Das wirft Fragen zum Datenschutz und zur Überwachung im öffentlichen Raum auf. Welches Unternehmen das ist, nennt das Material nicht.

RES

KI-Pionier Rich Sutton spricht über Kreativität und Entdeckung durch KI

Rich Sutton äußert sich zu Kreativität und Entdeckung im Kontext von KI. Inhaltliche Details aus dem Gespräch enthält das Material nicht. Der Beitrag verweist auf ein Video, dessen Inhalt hier nicht bewertet werden kann.

TOOLS & HANDS-ON

OS

DeepSeek-V4-Pro: Neues Open-Source-Textmodell mit über 4 Mio. Downloads

DeepSeek veröffentlicht V4-Pro als Open-Source-Modell für Textgenerierung und Konversation. Mit bereits 4 Millionen Downloads zählt es zu den meistgenutzten neuen Modellen auf HuggingFace.

OS

Google veröffentlicht DiffusionGemma-26B: Bild-zu-Text per Diffusion

Das Modell diffusiongemma-26B-A4B-it nimmt Bilder entgegen und erzeugt Text – ungewöhnlich, weil es dafür einen Diffusionsansatz statt klassischer Vorhersage nutzt. Downloads stehen noch aus.

OS

DiffusionGemma erzeugt Text 4-mal schneller als bisherige Gemma-Modelle

DeepMind bewirbt DiffusionGemma mit einem vierfachen Geschwindigkeitsvorteil bei der Textgenerierung. Das Modell nutzt einen Diffusionsprozess statt des üblichen Token-für-Token-Ansatzes.

PROD

GitHub Copilot CLI versteht jetzt Code-Struktur statt nur Text zu durchsuchen

Mit konfigurierten Language-Server-Protokollen (LSP) kann Copilot im Terminal echte Code-Intelligenz nutzen – statt wie bisher stupide nach Textmustern zu suchen. Das macht Vorschläge präziser.

OS

HelixDB: Graph-Datenbank auf Object-Storage mit eingebetteter Vektorsuche

HelixDB ist eine OLTP-Graphdatenbank, die direkt auf Object-Storage läuft und native Vektorsuche mitbringt – ein Jahr nach dem Erststart veröffentlichten die Entwickler eine neue Version.

PROD

NVIDIA-GPUs sichern Apples Private Cloud Compute nun auch auf Google Cloud

Apple weitet seinen Private Cloud Compute-Dienst auf Google Cloud aus und setzt dabei auf NVIDIA-GPUs mit Confidential Computing. Das soll sicherstellen, dass Server-Anfragen vertraulich bleiben.

AUSBLICK

Keine Termine gemeldet.