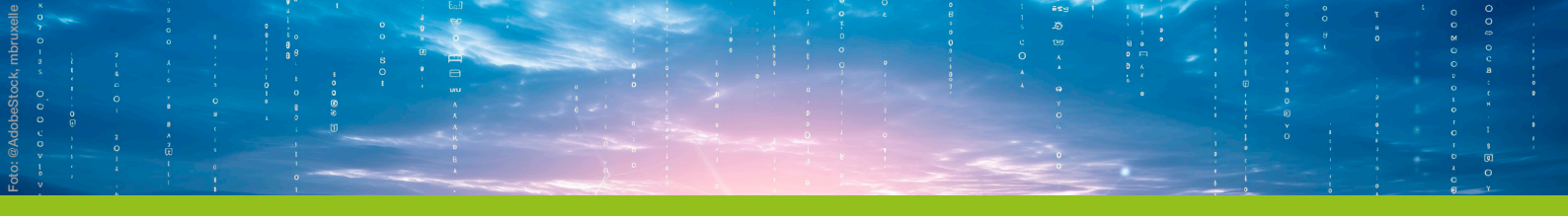
Foto: @AdobeStock, burararin

Künstliche Intelligenz in Unternehmen

KI-Modelle erfolgreich trainieren und einsetzen: Kriterien und Strategien zur Auswahl der richtigen IT-Infrastruktur

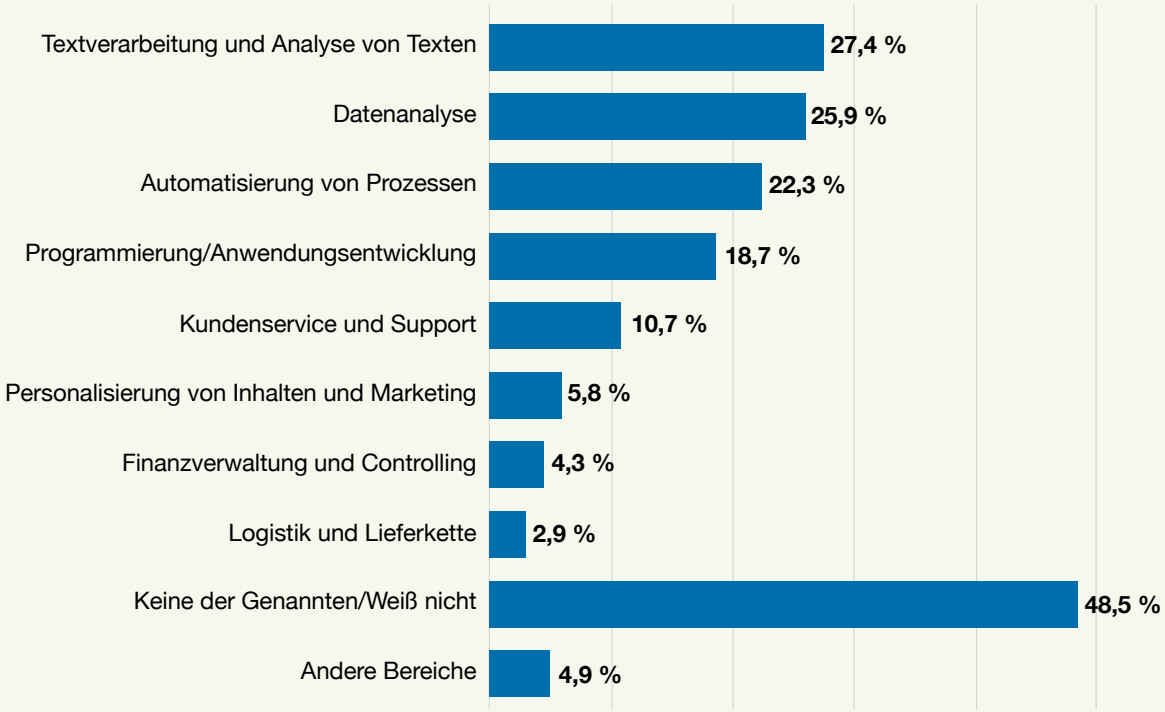
QCT | intel



Künstliche Intelligenz zählt zu den wichtigsten Zukunftstechnologien für Unternehmen. Das Training und der Einsatz von Machine-Learning-Algorithmen und anderen KI-Methoden stellt jedoch unterschiedliche Anforderungen an die IT-Infrastruktur – je nach Modell und Einsatzzweck. Dieses Whitepaper beleuchtet die grundlegenden Bereitstellungsmodelle für KI, erläutert die jeweiligen Vor- und Nachteile und zeigt, wie ein typischer Workflow zwischen verschiedenen Infrastrukturen aussehen kann. Hinzu kommen Beispiele für den Einsatz von KI sowie Empfehlungen für die Auswahl geeigneter Hard- und Software.

Mit dem Hype um ChatGPT hat auch in Deutschland die Diskussion um künstliche Intelligenz (KI) an Fahrt gewonnen. Rund zwei Drittel der deutschen Unternehmen halten sie laut einer Umfrage des Branchenverbands Bitkom für die wichtigste Zukunftstechnologie. Bei einer Umfrage des Meinungsforschungsinstituts Civey im Auftrag des Internetverbands eco gaben über 50 Prozent der Befragten an, KI im Unternehmen zu nutzen. Zu den häufigsten Anwendungszwecken zählen Textanalysen (27,4 Prozent), Datenanalysen (25,9 Prozent) und Prozessautomatisierungen (22,3 Prozent).

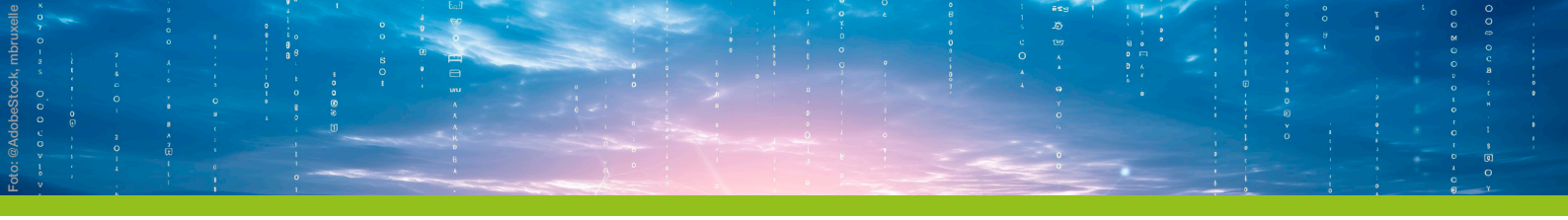
In welchen Bereichen setzen Sie bereits künstliche Intelligenz in Ihrem Unternehmen/bei Ihrem Arbeitgeber ein? (IT-Entscheider)



Mehrfachantwort möglich | Stat. Fehler Gesamtergebnis: 8,1% | Stichprobengröße: 504 | Befragungszeitraum: 15.03.24 – 31.03.24

Quelle: eco/Civey

Deutsche Firmen setzen KI vornehmlich für die Verarbeitung und Analyse von Texten, die Datenanalyse und die Prozessautomatisierung ein.

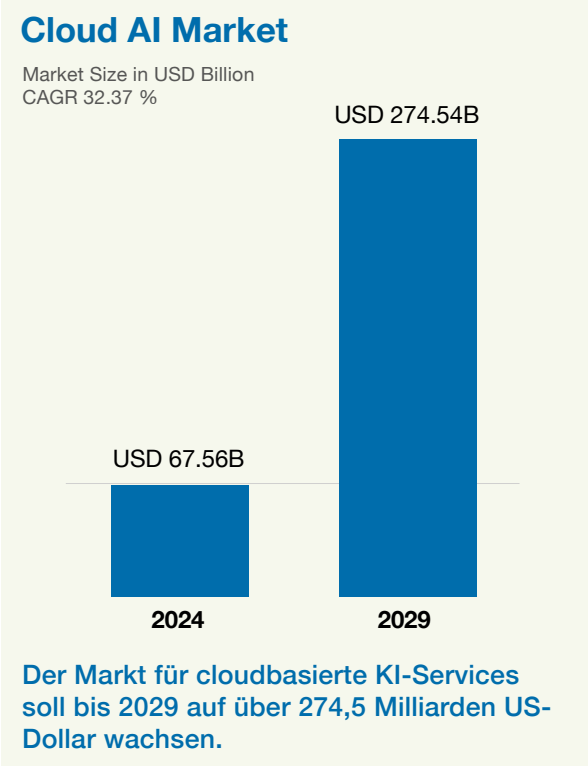


Bereitstellungsmodelle für KI

Das Training und die Anwendung von Machine-Learning-Algorithmen und anderen KI-Methoden stellt die IT-Infrastruktur vor unterschiedliche Herausforderungen. Es empfiehlt sich deshalb, vor der Investition in Hardware oder Services genau zu analysieren, welche Anforderungen der geplante Einsatzzweck mit sich bringt und welche IT-Infrastruktur sich dafür am besten eignet. Grundsätzlich existieren drei Bereitstellungsmodelle.

1. Cloud-Computing

Die Nachfrage nach KI-Services aus der Cloud soll in den kommenden Jahren stark zunehmen. Marktforscher prognostizieren für den globalen Cloud-KI-Markt bis 2029 ein durchschnittliches jährliches Wachstum von über 30 Prozent. Das Volumen wird von derzeit rund 67,5 Milliarden US-Dollar auf über 274 Milliarden US-Dollar steigen. Insbesondere die Entwicklung und das Training von Machine-Learning-Algorithmen und anderen KI-Methoden sind ressourcenintensiv und werden daher häufig in die Cloud ausgelagert. Aber auch wenn es um die Anwendung bestehender KI-Modelle, das sogenannte Inferencing, geht, erleichtern Cloud-Dienste den Einstieg. Microsoft bietet beispielsweise fast 30 verschiedene Services für KI und Machine Learning an. Die Einsatzgebiete reichen von der Mustererkennung über die Bot-Erstellung bis hin zu Computer Vision und Spracherkennung. Ähnlich umfangreich ist das Angebot an Machine-Learning- und KI-Services bei Amazon Web Services (AWS).



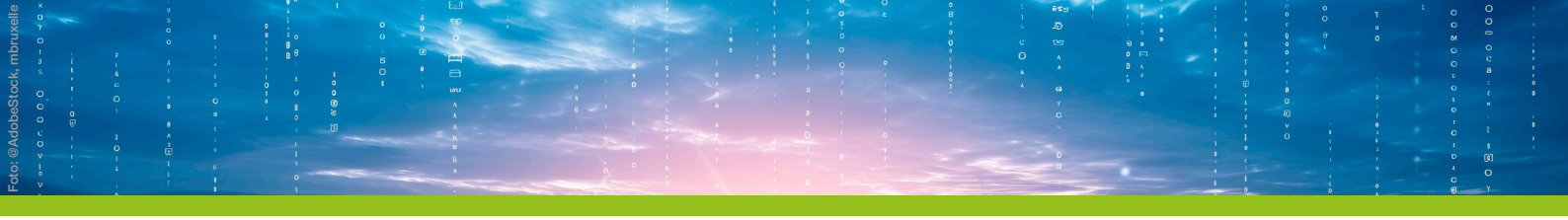
Der Markt für cloudbasierte KI-Services soll bis 2029 auf über 274,5 Milliarden US-Dollar wachsen.

Quelle: Mordor Intelligence

Auch in der Google Cloud können Anwender aus einer Vielzahl von KI-Services wählen. Vertex AI ermöglicht es beispielsweise, generative KI-Anwendungen herzustellen und anzupassen, Basismodelle mit eigenen Daten zu optimieren, Bilder zu erzeugen oder Software-Code zu generieren. Andere Module unterstützen das Erzeugen unternehmensspezifischer Suchmaschinen, die Dokumentenanalyse und die Modellentwicklung.

2. Lokales Rechenzentrum

Cloud-Ressourcen haben viele Vorteile: Sie sind sehr flexibel einsetzbar und nahezu unbegrenzt skalierbar. Zudem erfolgt ihre Abrechnung meistens nutzungsabhängig, der Einstieg gelingt also ohne hohe Investitionskosten. Während beispielsweise das Training von GPT-3 (Generative Pre-Trained Transformer 3, siehe auch „Die wichtigsten KI-Begriffe“) in der Cloud laut dem



KI-Architekten Lev Selector rund fünf Millionen US-Dollar gekostet hat, würde allein die Anschaffung entsprechender Hardware rund 30 Millionen US-Dollar verschlingen.

Der Vorteil der nutzungsabhängigen Abrechnung kann sich allerdings schnell ins Gegenteil verkehren, wenn der Ressourcenbedarf stark steigt. Der Datenwissenschaftler Hugo Debes hat ausgerechnet, wann die Nutzung von Chatbots auf LLM-Basis (Large Language Model) in der Cloud unwirtschaftlich wird. Seiner Kalkulation nach liegt die Grenze bei circa 8.000 Konversationen pro Tag.

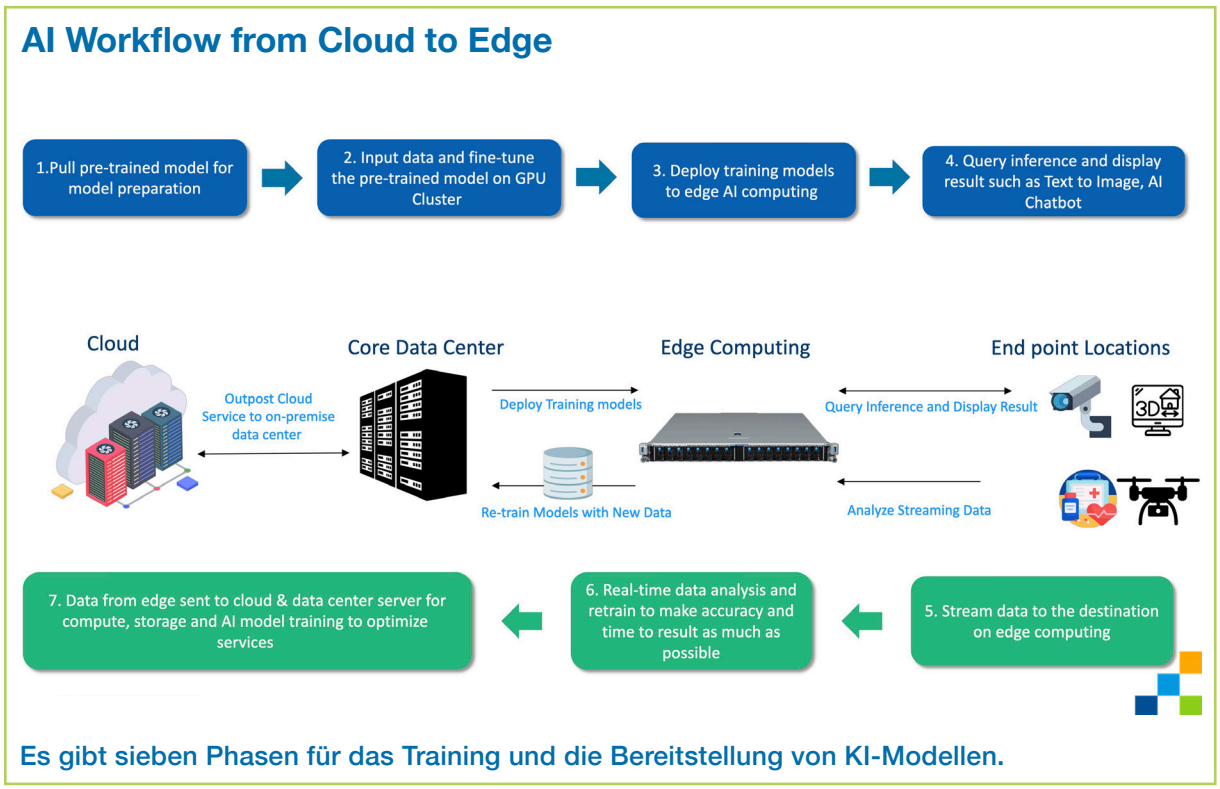
Schwerer als wirtschaftliche Überlegungen wiegen oft rechtliche Vorgaben wie die Datenschutzgrundverordnung (DSGVO) und Compliance-Anforderungen. Um Basismodelle an die unternehmensspezifischen Anforderungen anzupassen, müssen sie meist mit personenbezogenen Informationen und sensiblen Daten wie Finanzberichten oder Geschäftsgeheimnissen nachtrainiert werden, die nicht in fremde Hände gelangen dürfen. Public-Cloud-Umgebungen gewährleisten die notwendige Vertraulichkeit nicht und eignen sich deshalb nicht für diese Aufgabe. In stark regulierten Branchen wie Banken und Versicherungen, aber auch bei Behörden und anderen Einrichtungen der öffentlichen Hand kommt eine Public-Cloud-Nutzung ohnehin nicht infrage.

Für die Anwendung der KI-Modelle sind die Anforderungen an Rechenleistung und -kapazität meist deutlich geringer als für Entwicklung und Training, deshalb gelingt die Durchführung in der Regel auch im eigenen Rechenzentrum.

3. Edge

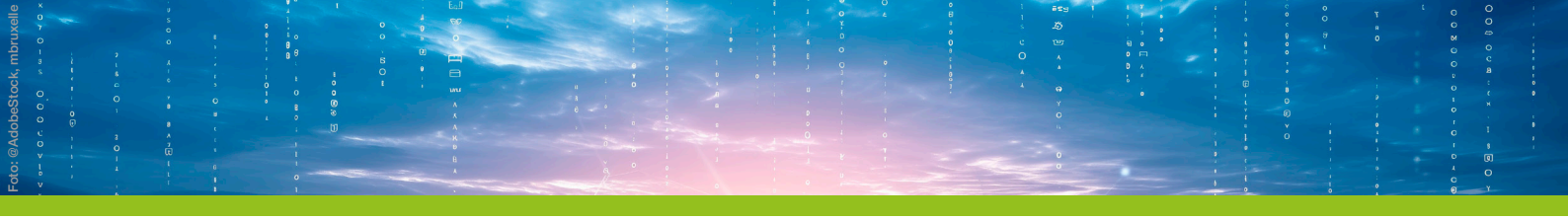
Die Zahl aktiver Geräte im Internet of Things (IoT) stieg laut IoT Analytics im Jahr 2023 um 16 Prozent auf fast 17 Milliarden IoT-Endpunkte. IoT-Geräte werden aber auch intelligenter. IoT Analytics schätzt, dass der Anteil smarter und KI-fähiger IoT-Endpunkte bis 2027 jährlich um 76 Prozent zunimmt.

Diese Entwicklung erhöht die Nachfrage nach Lösungen für KI direkt vor Ort, am sogenannten Edge. Sie kommen besonders dann infrage, wenn großen Datenmengen in Echtzeit verarbeitet werden müssen, etwa beim autonomen Fahren. Auch bei lückenhafter Netzabdeckung oder Funkverbindungen mit geringer Bandbreite ergibt die Verarbeitung vor Ort Sinn. Die Vorverarbeitung der Sensorinformationen direkt am Edge reduziert die ins Rechenzentrum oder die Cloud zu übertragende Datenmenge. Das senkt nicht nur die Kosten, sondern verringert auch die Anforderungen an Netzwerkgüte und -bandbreite.



Es gibt sieben Phasen für das Training und die Bereitstellung von KI-Modellen.

Quelle: QCT



KI-Workflow zwischen Cloud und Edge

Wie in der Abbildung dargestellt, lässt sich der Workflow zwischen Cloud, Data Center und Edge in sieben Phasen einteilen:

1. Entwicklung und Training eines Modells in der Cloud

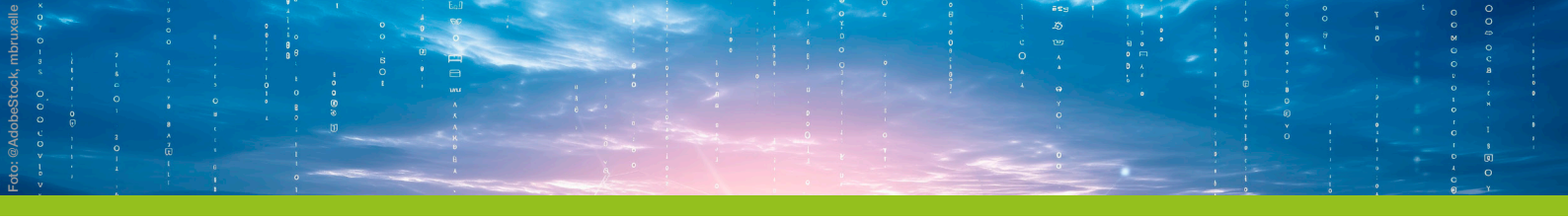
Die Entwicklung und das Training von großen Sprachmodellen wie GPT (Generative Pre-Trained Transformer) und anderen Basismodellen mit mehreren Hundert Milliarden von Parametern, aber auch von Computer-Vision-Algorithmen zur automatischen Bilderkennung erfolgen meist in einer Cloud-Umgebung. OpenAI, der Entwickler von ChatGPT, nutzt beispielsweise die Microsoft-Cloud Azure, um seine generativen Modelle zu entwickeln und zu trainieren. Google hingegen verwendet seine eigene Cloud-Infrastruktur für Modelle wie Gemini oder PaLM 2 (Pathways Language Model). Wer selbst Sprachmodelle in der Cloud trainieren möchte, kann auf Plattformen wie IBM WatsonX zurückgreifen oder Services wie Amazon SageMaker und Amazon-EC2-DL1 nutzen. Auch Intel Geti, eine Plattform zur Entwicklung von Computer-Vision-Lösungen, lässt sich auf virtuellen Instanzen in der Cloud installieren. Open-Source-Modelle wie Llama 2 von Meta oder Mistral 7B lassen sich ebenfalls in einer Cloud-Umgebung trainieren.

2. Finetuning des vortrainierten Modells auf einem GPU-Cluster im lokalen Rechenzentrum

Das eigene Rechenzentrum bietet mehr Kontrolle, wenn es um Data Governance und Compliance geht. Das Training beziehungsweise Finetuning von Modellen mit personenbezogenen oder geschäftskritischen Daten sollte deshalb besser On-Premise erfolgen. Der Datentransfer in die Modelle und zurück funktioniert schneller und unkomplizierter als bei der Cloud-Nutzung.
Für das Finetuning bieten sich grundsätzlich zwei Strategien an: Retrieval Augmented Generation (RAG) und Federated Learning. Bei RAG werden im Unternehmen vorhandene Dokumente wie technische Spezifikationen, Verträge oder Geschäftsberichte mit einer in der Cloud vortrainierten generativen KI verknüpft. Das Modell kann auf die zusätzlichen Informationen zugreifen und so bessere, unternehmensspezifische Antworten geben. Beim Federated Learning erfolgt das Training eines Modells auf mehrere Organisationen verteilt. Die Teilnehmer übermitteln ihre Ergebnisse regelmäßig an einen Aggregator („Orchestrator“), der sie zu einem konsolidierten Machine-Learning-Modell (Konsensmodell) zusammenfasst und das aktualisierte Modell zum weiteren Finetuning an die beteiligten Organisationen zurückgibt.

3. Bereitstellung des fertigen Modells am Edge

Das trainierte Modell landet nun entweder auf einem Edge-Server oder auf einem Endgerät. Server bieten mehr Rechenleistung und Speicherkapazität und empfehlen sich für anspruchsvolle Aufgaben mit großen Datenmengen sowie komplexe Analysen. Auch wenn die Daten mehrerer Geräte gleichzeitig in die Analyse einfließen, muss ein Server zum Einsatz kommen. Die Bereitstellung direkt auf dem Endgerät empfiehlt sich besonders dann, wenn die Datenverarbeitung in Echtzeit erfolgen soll und die Ergebnisse sofort vorliegen müssen, etwa in autonomen Fahrzeugen, Maschinen und IoT-Geräten. Für diese Aufgabe eignen sich Frameworks wie TinyML, die wenig Rechenleistung und Speicher benötigen.



4. Daten-Streaming am Edge

Je nach Einsatzmodell findet die Datenerfassung ganz oder teilweise getrennt von der KI-basierten Verarbeitung statt. Intelligente Überwachungskameras schaffen es beispielsweise, Personen, Fahrzeuge oder Tiere zu identifizieren und Gesichter zu erkennen. Diese vorverarbeiteten Informationen werden dann zur weiteren Analyse an den Edge-Server übertragen. Bei Kameras und Sensoren ohne intelligente Vorverarbeitung erfolgt die komplette Analyse auf dem Edge-Server.

5. Inferenzanfrage und Ergebnisdarstellung

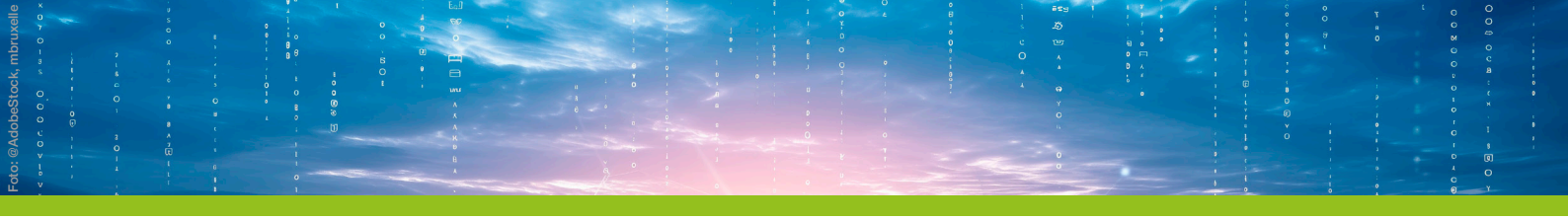
Das intelligente Endgerät oder der Edge-Server analysiert nun die Daten mithilfe des KI-Modells (Inferenz) und sendet das Ergebnis zurück. Die Art der Ergebnisdarstellung hängt vom jeweiligen Einsatzzweck ab. Bei Überwachungssystemen kann es beispielsweise zum Auslösen eines Alarms kommen, medizinische Geräte informieren unter Umständen Rettungskräfte oder einen Arzt, wenn sie eine lebensbedrohliche Situation feststellen. Produktionsanlagen reduzieren ihre Geschwindigkeit oder stoppen komplett, wenn die KI einen bevorstehenden Ausfall prognostiziert.

6. Übertragung der Daten ins Rechenzentrum und erneutes Training der Modelle

Machine-Learning-Modelle befinden sich in einem kontinuierlichen Entwicklungsprozess und erfordern fortlaufendes Training und Optimierung. Veränderungen der Rahmenbedingungen (neue Vorschriften, Prozesse oder Technologien) erfordern eine Anpassung der Modelle. Die am Edge gesammelten Daten sind deshalb wertvoll und sollten für das weitere Training und Finetuning des Modells ins Rechenzentrum wandern.

7. Datentransfer in die Cloud zur weiteren Optimierung der Modelle

Auch die in der Cloud trainierten Basismodelle lernen ständig dazu, wofür sie neue Daten benötigen. Sofern es sich nicht um personenbezogene oder andere sensible Informationen handelt, sollten die gewonnenen Erkenntnisse deshalb dem Cloud-Modell zur Verfügung stehen. Cloud-Ressourcen können außerdem für die Speicherung der gewonnenen Datenmengen genutzt werden, sofern eine starke Verschlüsselung gewährleistet ist.



Beispiele für den KI-Einsatz in Unternehmen

Laut der Digitalstrategie der Bundesregierung gehört künstliche Intelligenz zu den Schlüsseltechnologien einer innovativen Wirtschaft. Tatsächlich lässt sie sich schon heute in Firmen auf vielfältige Weise einsetzen. Hier einige Beispiele:

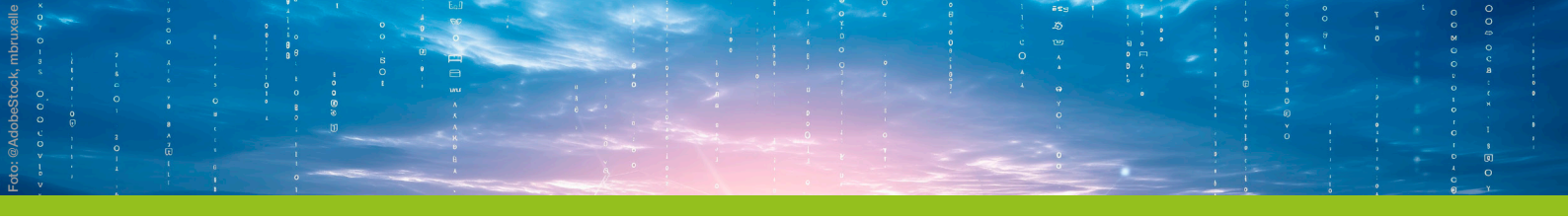
- **Optimierung von Industriedesigns:** KI kann helfen, die Konstruktion von Produkten und Maschinen zu verbessern, neue Designs zu entwickeln, Herstellungskosten zu senken und den Produktentwicklungsprozess zu beschleunigen.

- **Vorhersage von Trends:** Machine Learning und andere KI-Verfahren ermöglichen es, in großen Datenmengen Muster zu erkennen, Zusammenhänge in Millionen von Datenpunkten zu identifizieren, Vorhersagen zu treffen und Entscheidungen vorzubereiten. So lassen sich Trends schneller und zuverlässiger erkennen.

- **Prozessautomatisierung:** Unternehmen müssen Prozesse erkennen und verstehen, um Lücken und Engpässe zu identifizieren und die Produktivität der Angestellten zu verbessern. Dies lässt sich beispielsweise über die KI-gestützte visuelle Prozesserkennung (Visual Process Detection, VPD) erreichen. Sie erfasst, dokumentiert und analysiert Interaktionen zwischen Benutzern und Workflows in Echtzeit und identifiziert Prozesse, die automatisiert werden können.

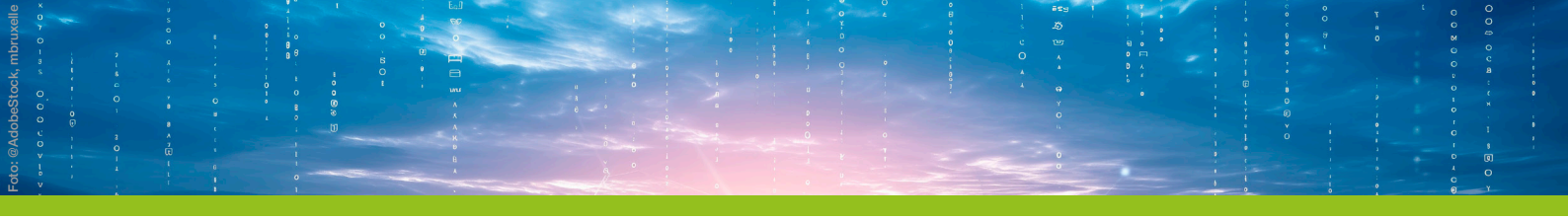
- **Identifizierung und Schutz personenbezogener Daten:** Daten enthalten oft Informationen zur Identifizierung von Personen. Werden diese Daten missbraucht, kann es zu Eingriffen in die Privatsphäre und die Selbstbestimmung bis hin zu Identitätsdiebstahl oder Diskriminierung kommen. Gesetze wie die DSGVO sollen diese Gefahren minimieren und die Privatsphäre der Menschen in der EU schützen. Diese Vorschriften schränken jedoch die Nutzung von Daten für betriebliche Zwecke ein. KI hilft dabei, personenbezogene Informationen in Datensätzen zu erkennen, um sie zu entfernen, zu anonymisieren oder zu verschlüsseln.

- **Vorhersage von Lieferzeiten:** Laut dem E-Commerce-Lieferkompass 2023 sind lange Lieferzeiten in 45 Prozent der Fälle der Grund dafür, dass Kunden einen Einkauf abbrechen. In Deutschland liegt die gerade noch akzeptierte Lieferzeit bei rund drei Tagen. Einer Umfrage von YouGov und ParcelLab zufolge bestellt ein Viertel der Online-Shopper nach einer negativen Erfahrung mit dem Versand nicht mehr bei einem Händler. Eine schnelle und pünktliche Lieferung stärkt die Kundenbindung. 46 Prozent der Kunden wünschen sich laut der YouGov-Studie zudem regelmäßige Updates zum Lieferstatus. Maschinelles Lernen und andere KI-Formen können dabei helfen, Lieferzeiten präziser vorherzusagen.



- **Erkennung von Anomalien:** Abweichungen vom Normalzustand weisen auf Fehler, Betrugsversuche oder Gesundheitsrisiken hin. Finanzdienstleister nutzen sie zur Betrugsprävention, im Gesundheitswesen helfen sie, medizinische Probleme früher zu erkennen – und in der Cybersicherheit unterstützen sie IT-Sicherheitsverantwortliche dabei, unrechtmäßige Zugriffe abzuwehren und erfolgreiche Eindringversuche schneller zu identifizieren. In der Fertigung kann die Erkennung von Anomalien dazu beitragen, den Betrieb und den Energieverbrauch zu optimieren oder die Ausfallzeiten von Anlagen zu minimieren.

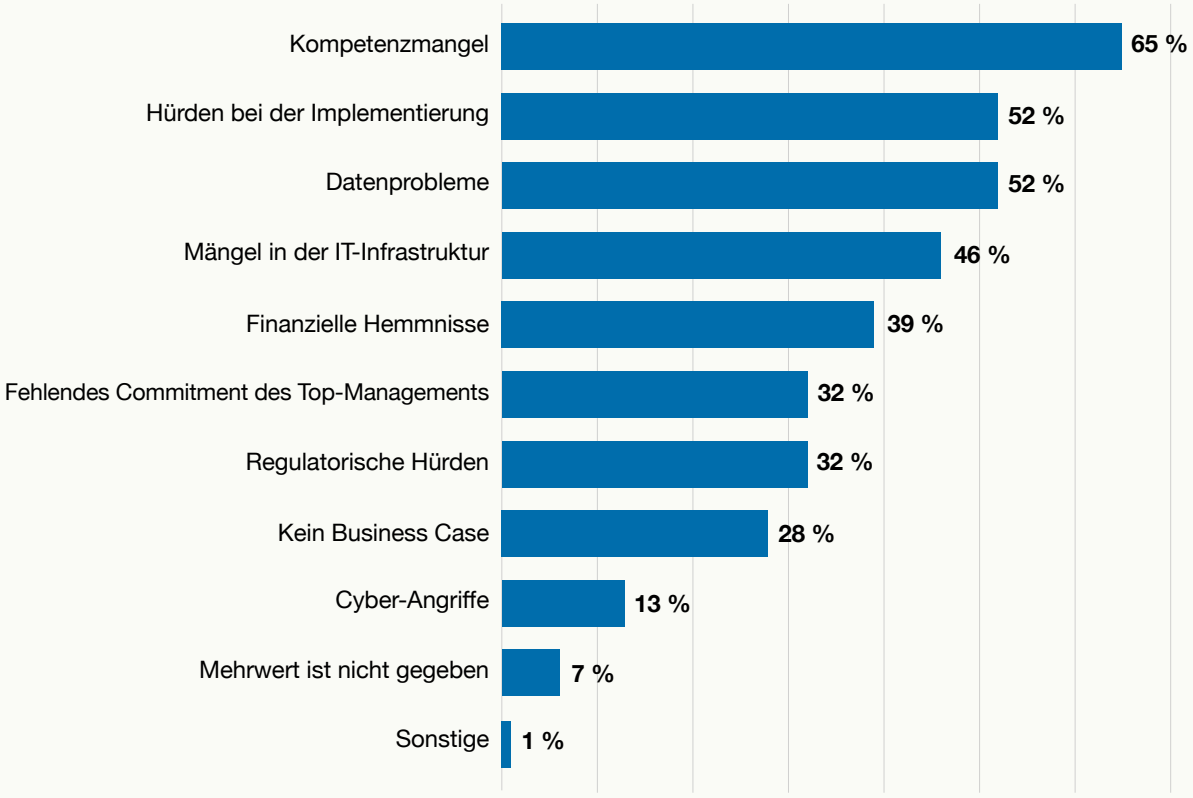
- **Betrugserkennung:** Laut Nilson Report lag der Verlust durch Kreditkartenbetrug im Jahr 2022 bei knapp 33,5 Milliarden US-Dollar. Bis 2028, so die Analysten, wird dieser Betrag auf fast 43,5 Milliarden US-Dollar ansteigen. Banken haben deshalb ein großes Interesse daran, betrügerische Kreditkartentransaktionen in Echtzeit zu erkennen. Machine-Learning-Modelle spielen dabei eine wichtige Rolle, indem sie betrügerische Transaktionen schneller und präziser erkennen als herkömmliche statistische Methoden.



Mit Referenz-Kits schneller zum Erfolg

Laut dem Report „State of AI in the Enterprise“ der Unternehmensberatung Deloitte hat rund die Hälfte der deutschen Unternehmen im Mittelstand Schwierigkeiten, KI-Technologien zu implementieren und die notwendigen Daten für das Training der Modelle bereitzustellen.

Hemmnisse der KI im Mittelstand (Mehrfachnennungen möglich)



Neben fehlender Kompetenz behindern vor allem Hürden bei der Implementierung und Datenprobleme die Einführung von KI im Mittelstand.

Quelle: Deloitte

Diese Hürden lassen sich durch den Einsatz spezieller Bibliotheken senken, wie sie etwa der Hersteller Intel in Zusammenarbeit mit der Technologieberatung Accenture für eine Vielzahl von Anwendungsfällen entwickelt hat. Die Intel AI Reference Kit Library basiert auf Technologien wie OneAPI, OpenVINO (Open Visual Inference & Neural Network Optimization), Intel Advanced Matrix Extensions (Intel AMX), SYCL und Neural Compressor. Jedes Referenz-Kit enthält ein auf die jeweilige Branche zugeschnittenes vortrainiertes KI-Modell sowie einen angepassten Workflow für die Machine-Learning-Pipeline. Derzeit stehen mehr als 30 Referenz-Kits auf GitHub zur Verfügung. Um Geschäftsprozesse vollständig abzubilden, lassen sich mehrere Kits kombinieren.

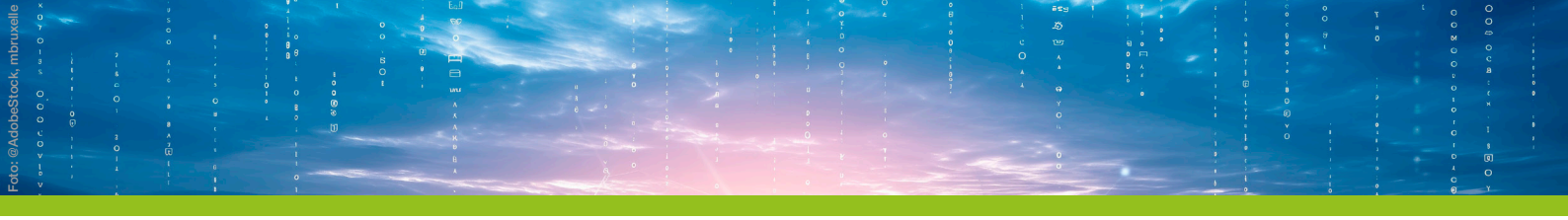


Foto: @AdobeStock, mbruxelle

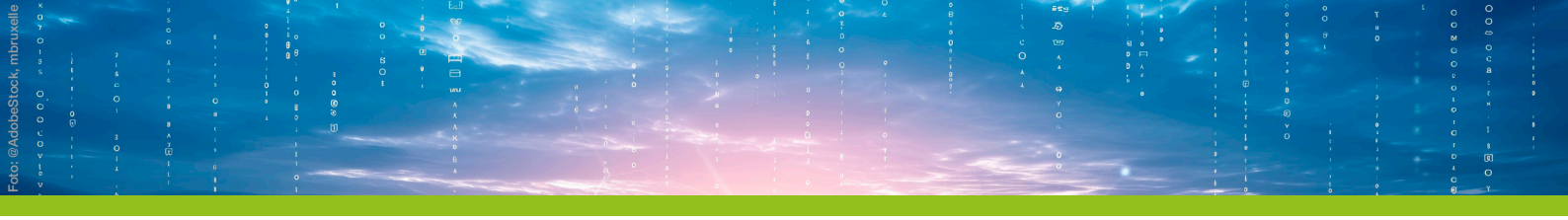
Das Referenz-Kit für die Designoptimierung basiert beispielsweise auf einem vortrainierten Generative Adversarial Network (GAN), mit dem sich realistische Designentwürfe entwickeln lassen. Hinzu kommen tiefe neuronale Netze für die Feature-Erkennung und -Extraktion, ein Trainingsdatensatz sowie für das Training notwendige Bibliotheken und Softwarekomponenten (Intel Extension for PyTorch, Intel Neural Compressor). Mit Intel-oneAPI-Komponenten lässt es sich mit wenigen Modifikationen an die eigenen Anforderungen anpassen.

Das Referenz-Kit zur Vorhersage von Krankheiten nutzt die KI-gestützte Verarbeitung natürlicher Sprache (Natural Language Processing, NLP), um Diagnosen und Gesundheitsrisiken in Echtzeit aus Befunden zu extrahieren. Das Kit basiert auf dem Modell ClinicalBERT (Bidirectional Encoder Representations from Transformers), das mit fast 5.000 Symptombeschreibungen und den darauf basierenden Diagnosen trainiert wurde. Laut Intel gelingt so die Erkennung von Krankheiten um fast 70 Prozent häufiger.

Mit dem Referenz-Kit für visuelle Prozesserkennung schaffen es Unternehmen, webbasierte Prozesse zu optimieren. Das Modell verwendet ein vortrainiertes gefaltetes Netz (Convolutional Neural Network, CNN), um interaktive Webseitenelemente wie Buttons, Links oder Eingabefelder zu identifizieren. Das Kit beschleunigt nicht nur die Erkennung der Elemente um bis zu 150 Prozent, sondern agiert auch besonders ressourceneffizient. Dank reduzierter Datenmenge und besonders effizienter Algorithmen lässt es sich auch auf Edge-Geräten mit geringer Prozessorleistung und wenig Arbeitsspeicher ausführen.

Das Referenz-Kit für die Anonymisierung personenbezogener Daten hilft dabei, sensible Informationen in großen Datenmengen zu finden und diese zu anonymisieren. Ein vortrainiertes BERT-Modell erkennt relevante Einträge in Texten. Die Maskierung der identifizierten Daten übernimmt anschließend ein Recurrent Neural Network (RNN), das in der Lage ist, realistisch wirkende Zufallsnamen zu erzeugen. Mithilfe des Referenz-Kits können Anwender mehrere Terabyte an Daten auf einer einzigen Workstation analysieren. Das Modell lässt sich außerdem nahtlos von lokalen Rechnern in eine Cloud-Umgebung erweitern, ohne dass der Code geändert werden muss.

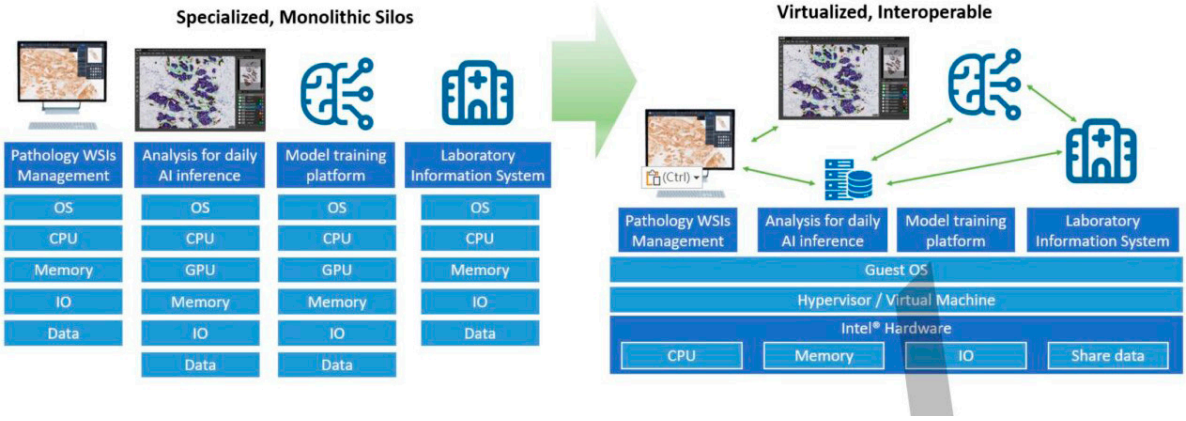

Foto: @AdobeStock, Mediaparts



Künstliche Intelligenz in der Medizin – ein Praxisbeispiel

Digitale Gesundheitslösungen mit KI-Unterstützung verändern das Gesundheitswesen, indem sie die Ergebnisse für die Patienten verbessern, den Zugang erleichtern und die Effizienz steigern. Innovationen wie Telemedizin, mobile Gesundheitsanwendungen und KI-gestützte Diagnostik ermöglichen präzise Diagnosen und individuelle Behandlungen. In der Pathologie verringert die Digitalisierung von Objektträgern in hochauflösende Bilder, die am Computer betrachtet werden können, nicht nur den Platzbedarf im Lager, sondern verbessert auch die Geschwindigkeit und Genauigkeit komplexer Diagnosen.

JelloX Biotech ist ein Start-up-Unternehmen, das sich auf digitale 3D-Pathologiebilder mit KI-Analyse spezialisiert hat. Obwohl KI die Effizienz der Analyse verbessern kann, löst sie nicht das grundlegende Problem der niedrigen Rate an Biopsieproben. Die 3D-Pathologie ermöglicht es Pathologen, wesentlich mehr Daten zu sammeln und so eine umfassendere und genauere Diagnose zu erstellen. Mit mindestens hundertmal mehr Daten im Vergleich zu herkömmlichen 2D-H&E- und IHC-Objektträgern gelingt es dann effizienter, die Rechen- und Analyseleistung der KI zu nutzen.

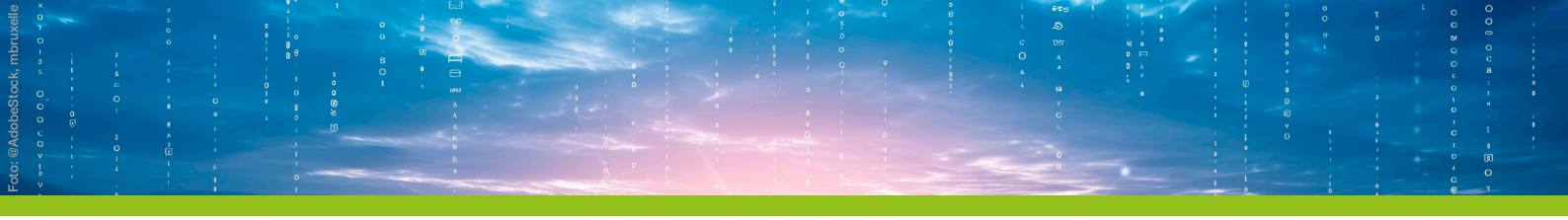


JelloX hat eine konsolidierte Lösung für die KI-basierte Analyse von pathologischen Bildern entwickelt.

Dank der Intel Advanced Matrix Extensions (Intel AMX) lassen sich KI-Modelle direkt auf dem Hauptprozessor (CPU) trainieren und anwenden, der teure und zeitaufwendige Transfer in eine GPU-Infrastruktur (Graphic Processing Unit) und zurück entfällt. Die großzügige Ausstattung mit Arbeitsspeicher ermöglicht es, größere Datenmengen auf einmal zu verarbeiten, als dies in GPU-Umgebungen möglich wäre. Das beschleunigt und erleichtert das Training mit hochkomplexen Datensätzen. Mit den skalierbaren Intel-Xeon-Prozessoren entsteht außerdem eine hochsichere Umgebung für die Verarbeitung sensibler Daten.

Mithilfe der Intel Software Guard Extensions (Intel SGX) beschränken Entwickler die Ausführung von Code auf einen geschützten Bereich der CPU, die sogenannte Enklave. Die Entschlüsselung der Daten erfolgt nur in dieser geschützten Laufzeitumgebung (Intel Trusted Execution Technology, TXT), was sie vor neugierigen Blicken schützt.

Für die Entwicklung seiner Lösung verwendete JelloX die QCT DevCloud, eine in Zusammenarbeit mit Intel entwickelte Plattform für die Evaluierung und das Testen von HPC- (High-Performance Computing) und KI-Workloads. Vor



der Kaufentscheidung für eine bestimmte Rechenarchitektur können Entwickler auf der QCT DevCloud ihre Workloads testen und die beste Kombination für ihre Anforderungen finden. Die Plattform ermöglicht Entwicklern den Fernzugriff auf eine cloudnative oder Bare-Metal-Umgebung. Aktuell unterstützt die DevCloud die fünfte Generation der skalierbaren Intel-Xeon-Prozessoren, in Zukunft wird auch die sechste Generation verfügbar sein. Die GPUs der Flex-Serie (Flex 140, Flex 170) werden ebenfalls unterstützt. Als Server stehen unter anderem folgende Modelle zur Verfügung:

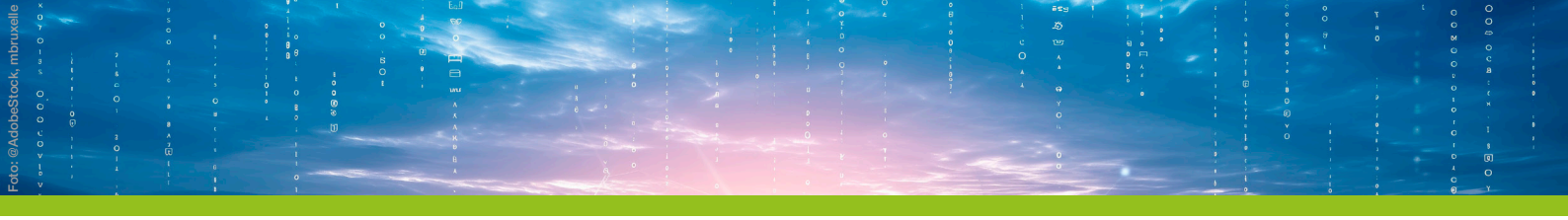
QuantaGrid D54X-1U: Der eine Höheneinheit (HE) messende Server bietet durch sein modulares Design hohe Flexibilität. Er lässt sich mit skalierbaren Intel-Xeon-Prozessoren der vierten und fünften Generation bis zu einer thermischen Verlustleistung von 350 Watt (385 Watt mit Flüssigkeitskühlung) betreiben und ist mit 16 NVME-SSDs im E1.S-Design oder zwölf 2,5 NVME-SSDs ausgestattet. Es werden bis zu acht TB DDR5-Arbeitsspeicher unterstützt. Das System eignet sich für Cloud-, Enterprise-, KI-, HPC-, Netzwerk-, Sicherheits- und IoT-Workloads.

QuantaGrid D54Q-2U: Dieser zwei HE hohe Server für zwei Xeon-Prozessoren ist für die Beschleunigung von KI-Aufgaben optimiert. Er unterstützt bis zu 24 NVME-SSDs im U.2- oder E1.S-Format. Das System eignet sich für Cloud-, Enterprise-, KI-, HPC-, Netzwerk-, Sicherheits- und Speicher-Workloads.

QuantaGrid D54U-3U: Dieser Server bietet im drei Höheneinheiten messenden Rackformat zusätzlich zu den zwei Intel-Xeon-Prozessoren der vierten und fünften Generation verschiedene GPU-Konfigurationen. Er eignet sich besonders für das Training und die Anwendung großer KI-Modelle.

Fazit: KI benötigt nahtlose und flexible Infrastrukturen

Die Einsatzmöglichkeiten künstlicher Intelligenz sind nahezu unbegrenzt und bislang nur in Ansätzen realisiert. Unternehmen sollten so schnell wie möglich ihre Prozesse und Geschäftsmodelle analysieren und prüfen, wie und in welchem Umfang KI-Unterstützung für mehr Wachstum und Effizienz sorgen kann. Dabei ist es entscheidend, auf eine möglichst flexible Plattform zu setzen, die Cloud-Umgebungen, lokale Rechenzentren und Edge-Computing nahtlos vernetzt. Nur so lassen sich die unterschiedlichen Anforderungen beim Training und der Anwendung von KI schnell und kostengünstig umsetzen.



Die wichtigsten KI-Begriffe

BERT (Bidirectional Encoder Representations from Transformers): Bei BERT handelt es sich um ein von Google entwickeltes Sprachmodell, das auf der Transformer-Architektur basiert. Es wurde 2018 publiziert und stellt einen großen Fortschritt im Verständnis natürlicher Sprache dar.

Convolutional Neural Networks (CNN): CNN bestehen aus einer oder mehreren Faltungsschichten (Convolutional Layer) und einer Ergebnisschicht (Pooling Layer). Neuronen reagieren nur auf lokale Signale aus der vorhergehenden Schicht. So gelingt es beispielsweise, Kanten in einem Bild schnell und zuverlässig zu erkennen. CNNs arbeiten funktionell ähnlich wie die Sehrinde im menschlichen Gehirn und eignen sich für Aufgaben des maschinellen Sehens (Computer Vision).

Basismodelle: Sogenannte Basismodelle (Foundation Models) werden auf einer breiten Datenbasis trainiert und unterstützen viele Anwendungsfälle. Sie lassen sich mithilfe von RAG relativ einfach an spezifische Anforderungen anpassen.

Inferenz: Der Prozess, bei dem ein trainiertes KI-Modell Daten erhält und daraus Schlussfolgerungen zieht. Dass das auch mit neuen Daten funktioniert, die nicht Teil des Trainings waren, macht die „Magie“ von KI aus.

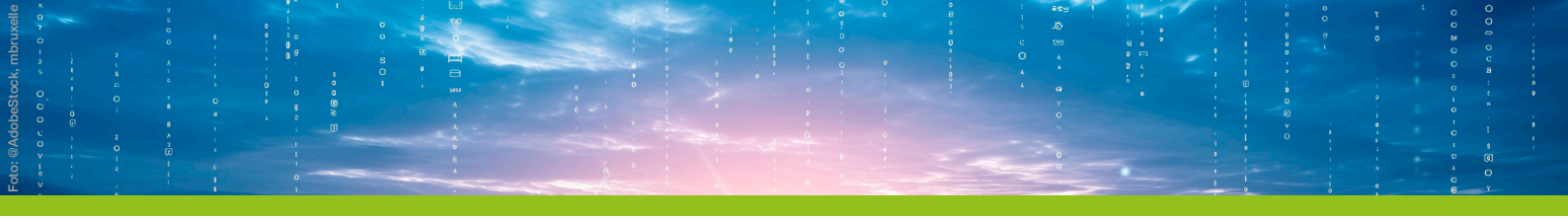
Generative Adversarial Networks (GAN): Ein GAN besteht aus zwei neuronalen Netzen, dem „Generator“ und dem „Diskriminator“. Das Generator-Netz erzeugt aus den Ausgangsdaten Varianten, die dem Original möglichst ähnlich sind. Das Diskriminator-Netz versucht, diese „Fälschungen“ vom Originaldatensatz zu unterscheiden. Während des Trainings liefert der Generator immer bessere Ergebnisse. GANs eignen sich unter anderem zur Erzeugung von realistisch wirkenden Bildern, Videos oder Audiodateien.

Generative Pre-Trained Transformer (GPT): Transformer-Netze gehören zur Klasse der großen Sprachmodelle. Sie sind in der Lage, Wortsequenzen oder Sätze zu erfassen, deren Bedeutung kontextabhängig zu extrahieren und eine auf den Input (Prompt) abgestimmte Antwort zu liefern. Die aktuell bekannteste Anwendung ist ChatGPT.

Large Language Model (LLM): Große Sprachmodelle sind neuronale Netze, die natürliche Sprache verarbeiten und generieren. Dazu analysieren sie in der Trainingsphase die statistische Wahrscheinlichkeit, mit der bestimmte Wörter oder Wortbestandteile aufeinander folgen. Das so entwickelte Modell wird durch menschliches Feedback weiter trainiert (Supervised Learning). LLMs schaffen es, auf eine Anfrage hin eine (mehr oder weniger) sinnvolle, grammatikalisch korrekte Antwort zu generieren.

Recurrent Neural Network (RNN): Während in sogenannten Feedforward-Netzen Informationen nur linear von einer Neuronenschicht zur nächsten wandern, kommunizieren RNNs in beide Richtungen. Dadurch können sie sich an zurückliegende Ereignisse „erinnern“ und zeitliche sowie logische Korrelationen herstellen. Einsatzgebiete: Erkennung von Handschriften, Sprache und Anomalien.

Retrieval Augmented Generation (RAG): LLMs und andere Basismodelle werden mit großen Datenmengen trainiert. In der Regel handelt es sich dabei um öffentlich zugängliche Informationen, etwa die Inhalte von Webseiten oder von Wikipedia. Mit RAG lassen sich solche Modelle auf spezifische Fragestellungen hin optimieren. Das Modell wird dazu mit internen Wissensdatenbanken wie zum Beispiel technischen Produktbeschreibungen oder Servicehandbüchern verknüpft. Bei einer Abfrage greift das LLM auf dieses Wissen zurück, was die Zuverlässigkeit und Genauigkeit der Antworten erhöht.



Über QCT

Quanta Cloud Technology (QCT) ist ein globaler Anbieter von Rechenzentrumslösungen. Das Unternehmen kombiniert die Effizienz von Hyperscale-Hardware mit Infrastruktur-Software von verschiedenen Branchenführern, um die Design- und Betriebsherausforderungen von Next-Generation-Rechenzentren zu lösen. QCT erbringt Dienstleistungen für Cloud Service Provider, Telekommunikationsunternehmen und Unternehmen, die öffentliche, hybride und private Clouds betreiben.

Zu den Produktlinien zählen hyperkonvergente und softwaredefinierte Rechenzentrumslösungen sowie Server, Speicher, Switches und integrierte Racks, ergänzt durch ein Ökosystem aus verschiedenen Komponenten- und Software-Partnern. QCT entwickelt, fertigt, integriert und erbringt Dienstleistungen für wegweisende Lösungen, die über das eigene globale Netzwerk angeboten werden. Die Muttergesellschaft von QCT ist Quanta Computer Inc., ein Fortune-Global-500-Unternehmen.

Quanta Cloud Technology
Germany GmbH
Rurbenden 48
52353 Düren

TEL: +49-2421-3863400
Fax: +49-2421-3863899

Dieses Whitepaper wurde erstellt von der eMedia GmbH, einer Tochtergesellschaft der Heise Media GmbH & Co. KG