Enterprise KI

Generative KI und domänenspezifische Modelle für Unternehmen

Optimieren Sie Training und Bereitstellung mit zweckspezifischer Intel® KI-Hardware und -Software zur Transformation Ihres Unternehmens

intel®

Inhalt

Warum Intel als Partner wählen?

Wir bei Intel haben die Möglichkeit, das Leben und die Ergebnisse für alle und für jedes Unternehmen auf diesem Planeten zu verbessern.

Aber wir tun dies nicht alleine!

Gemeinsam mit unseren Partnern schaffen wir echten Mehrwert für unsere Kunden, indem wir KI überall verfügbar machen und die Risiken bei der Bereitstellung minimieren

Wenn Sie mit Intel zusammenarbeiten, profitieren Sie von einem vollständigen KI-Technologieumfeld

Unser breitgefächertes Portfolio an KI-Technologien und Partnerschaften mit Hardware-, Software- und Systemintegratoren, die alle eng zusammenarbeiten, schaffen reale Lösungen, die differenzierte Geschäftsergebnisse für Branchen, Unternehmen und Gemeinden liefern.

Wir helfen Ihnen beim Wachstum Ihres Unternehmens.

Begleiten Sie uns auf dem Weg, KI überall verfügbar zu machen

Mit Intel® KI-Lösungen Mehrwert für Kunden schaffen

Der Ansatz von Intel ermöglicht es einem breiten, offenen Technologieumfeld von KI-Akteuren, Lösungen anzubieten, die unternehmensspezifische Anforderungen an generative KI erfüllen.

NAVER

Die Entwicklung eines leistungsstarken Large Language Model (LLM) zur globalen Bereitstellung von fortgeschrittenen KI-Diensten – in der Cloud bis hin zu Geräten. NAVER hat die grundlegenden Leistungsmerkmale von Intel® Gaudi® hinsichtlich der Ausführung von Rechenoperationen für groß angelegte Transformatormodelle mit hervorragender Leistung pro Watt bestätigt.

BOSCH

Zur Erkundung weiterer Möglichkeiten für die intelligente Fertigung, einschließlich fundamentaler Modelle, die synthetische Datenmengen von Fertigungsanomalien generieren, um robuste, gleichmäßig verteilte Trainingssätze bereitzustellen (z. B. automatisierte optische Inspektion).

IBM

Verwendet Intel® Xeon® Prozessoren der 5. Generation für seinen watsonx.data™ Datenspeicher und arbeitet eng mit Intel zusammen, um die watsonx™ Plattform für Intel® Gaudi® Beschleuniger zu validieren.

OLA

Das Unternehmen hat sein erstes indisches grundlegendes Modell mit generativen Funktionen in zehn Sprachen vorab trainiert und feinabgestimmt, sodass es im Vergleich zu marktüblichen Lösungen ein branchenführendes Preis-Leistungs-Verhältnis bietet. Krutrim führt derzeit ein Vorab-Training eines größeren grundlegenden Modells auf einem Intel® Gaudi® 2 Cluster durch.

seekr

Ein führender Anbieter vertrauenswürdiger KI führt Produktions-Workloads auf Intel® Gaudi® 2, Intel® Data Center GPU Max Series und Intel® Xeon® Prozessoren in der Intel® Tiber™ Developer Cloud zur Unterstützung der LLM-Entwicklung und der Produktionsbereitstellung aus.

IFF

Global führender Anbieter von Lebensmitteln, Getränken, Duftstoffen und Biowissenschaften wird generative KI und die Technik für digitale Zwillinge nutzen, um einen integrierten digitalen Biologie-Workflow für fortschrittliches Enzymdesign und die Optimierung von Fermentationsprozessen zu etablieren.

airtel

Airtel nutzt die Leistungsfähigkeit der hochmodernen Intel® Technologie und plant, seine umfangreichen Telekommunikationsdaten zu nutzen, um seine KI-Funktionen zu verbessern und das Erlebnis für seine Kunden zu optimieren. Die Bereitstellungen stehen im Einklang mit Airtels Bestreben, an der Spitze der technologischen Innovation zu bleiben und neue Einnahmequellen in einer sich schnell entwickelnden digitalen Landschaft zu generieren.

Infosys

Ein weltweit führender Anbieter von digitalen Services und Beratung der nächsten Generation, kündigte eine strategische Zusammenarbeit an. Ziel ist es, Intel® Technik wie Intel® Xeon® Prozessoren der 4. und 5. Generation, Intel® Gaudi 2 KI-Beschleuniger und Intel® Core™ Ultra Prozessoren in Infosys Topaz einzubringen – eine Reihe von Services, Lösungen und Plattformen mit KI-first-Ansatz, die mit generativer KI-Technik den Unternehmenswert steigern.

Zusammenarbeit des Technologieumfelds zur Entwicklung offener Plattform für Enterprise KI

Enterprise KI – Werteversprechen

Transformation Ihres Unternehmens mit Enterprise KI

In dem hyperkompetitiven Umfeld von heute **sind Unternehmen, die KI nutzen, weiter vorne mit dabei.**

Unternehmen aus allen Branchen durchleuchten jeden Betriebsaspekt, um zu verstehen, wie KI Arbeitsabläufe verbessern oder sogar automatisieren kann.

Die Einbindung von KI in die Unternehmensstruktur ist eine einzigartige Kompetenz von Intel.

Von KI-PCs, die die Produktivität transformieren, bis hin zu jahrelanger Fachkenntnis darüber, welche Anwendungsfälle den größten Nutzen erbringen: Intel® ist Ihr vertrauenswürdiger Partner, um KI überall, sicher und verantwortungsvoll einzusetzen.

Es wird erwartet, dass generative KI (GenAI)-Innovationen von Unternehmen aller Größenordnungen schneller übernommen werden als das Internet-Zeitalter, das mobile Zeitalter oder das Cloud-Zeitalter.

Die nächste Welle von KI-Plattformen wird diese faszinierenden Realitäten auf erschwingliche und flexible Weise aufgreifen.

Ein neuer Denkansatz für Ihre Enterprise KI ist jetzt gefragt.

Dieses Enablement-Paket hilft Ihnen zu verstehen, wie Unternehmen in verschiedenen Märkten einen erheblichen Nutzen aus generativer KI ziehen können, insbesondere aus domänenspezifischen Modellen – für einen langfristigen Erfolg.

Was ist generative KI und was sind Large Language Models?

Generative KI (GenAI) ist ein Teilbereich der KI, dessen Schwerpunkt auf der Erstellung neuer, origineller Inhalte liegt.

Er umfasst das Training und die Bereitstellung von KI-Modellen, um Daten wie Bilder, Text oder Audio zu generieren, die Beispielen aus der Trainingsdatenmenge sehr ähnlich sind.

GenAI-Algorithmen verwenden fortschrittliche Technik wie Deep Learning und neuronale Netzwerke, um realistische und kohärente Ergebnisse zu erzeugen, die Anwendungen wie Bildsynthese, Textgenerierung und sogar kreative Kunstwerke ermöglichen.

Large Language Model (LLM) ist ein spezifisches Modell für die Verarbeitung natürlicher Sprache, das tiefe neuronale Netzwerke verwendet, um Text zu verarbeiten und zu generieren. LLMs werden mit riesigen Mengen an Textdaten trainiert und sind darauf ausgelegt, kohärente und bedeutsame Ergebnisse zu generieren.

Weitere Informationen

WEITERLESEN

Nutzen Sie die Vorteile generativer KI

Wie werden Unternehmen GenAI verwenden?

Konsumgüter und Einzelhandel

- Virtuelle Umkleideräume
- Lieferung und Installation
- Unterstützung bei der Produktsuche in Geschäften
- Bedarfsvorhersage und Bestandsplanung
- Neuartige Produktdesigns

Gesundheitswesen und Medizin

- Unterstützung für vielbeschäftigte Mitarbeiter
- Transkription und Zusammenfassung medizinischer Notizen
- Chatbots zur Beantwortung medizinischer Fragen
- Vorausschauende Analysen als Grundlage für Diagnose und Behandlung

Produzierende Industrie

- Experten-Copilot für Techniker
- Gesprächsbezogene Interaktionen mit Maschinen
- Vorschriftsmäßiger und proaktiver Außendienst
- Fehlerbehebung natürlicher Sprache
- Garantiestatus und Dokumentation
- Verstehen von Prozessengpässen, Entwickeln von Wiederherstellungs-strategien

Medien und Unterhaltung

- Intelligente Suche, maßgeschneiderte Content Discovery
- Headline und Copy Development
- Echtzeit-Feedback zur Content-Qualität
- Personalisierte Wiedergabelisten, News Digests, Empfehlungen
- Interaktives Storytelling mittels Zuschauerentscheidungen
- Zielgerichtete Angebote, Abonnementpläne

Finanzdienst-leistungen

- Aufspüren von Trading-Signalen, Alarmierung von Tradern bei gefährdeten Positionen
- Beschleunigung von Underwriting-Entscheidungen
- Optimierung und Neuaufbau älterer Systeme
- Reverse-Engineering von Banking- und Versicherungsmodellen
- Überwachung von potenziellen Finanzstraftaten und Betrug
- Automatisierung der Datenerfassung für die Einhaltung gesetzlicher Vorschriften
- Gewinnung von Erkenntnissen aus Unternehmens-informationen

Quelle: Zusammengestellt von MIT Technology Review Insights basierend auf Daten aus „Retail in the Age of Generative AI“,9 „The Great Unlock: Large Language Models in Manufacturing“,10 „Generative AI Is Everything Everywhere, All at Once“ und „Large Language Models in Media & Entertainment“,12 Databricks, April–Juni 2023.

Anwendungsfälle für generative KI und Large Language Models

Chatbots und virtuelle Assistenten

Kundenbetreuung

Code-Generierung und Debugging von LLMs

Trainiert mit Unternehmensdokumenten

Stimmungsanalyse

Bewerten der Kundenzufriedenheit

Textklassifizierung und Clustering

Kategorisierung großer Datenmengen zur Identifizierung von Trends

Sprache Übersetzung

Umsetzung von Unternehmenswebseiten in andere Sprachen

Zusammenfassung und Umschreibung

Besprechungsnotizen zusammengefasst

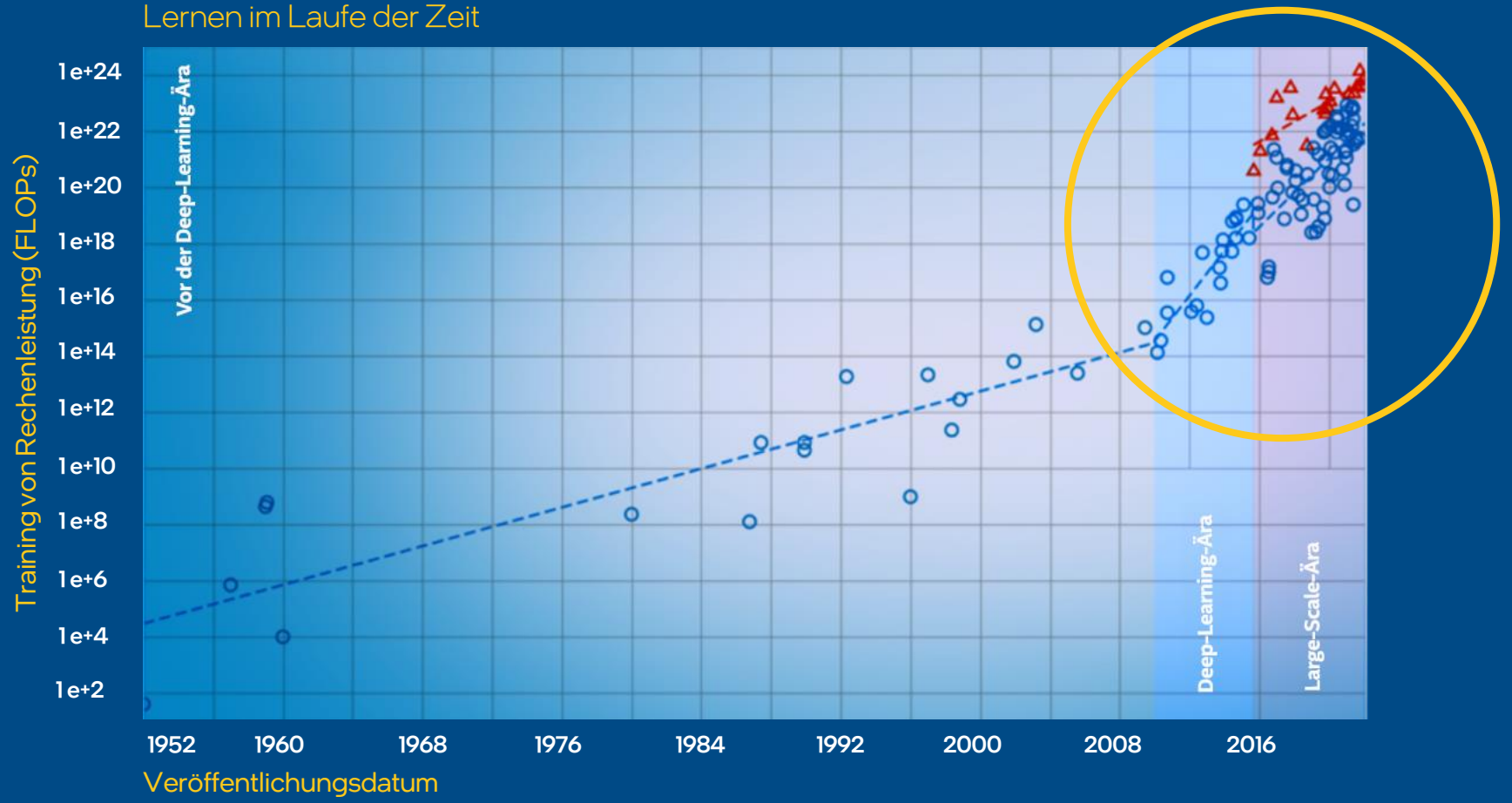
Content-, Bild-, Videogenerierung

Erste Entwürfe von E-Mails, Ideengenerierung, Marketing-Grafiken, kurzes Video

Quelle: A Compact Guide to Large Language Models, Databricks, 2023.

Mit der Größe der Modelle wächst auch die Rechenleistung

Training von Rechenleistung (FLOPs) von Milestone-Systemen für maschinelles Lernen im Laufe der Zeit



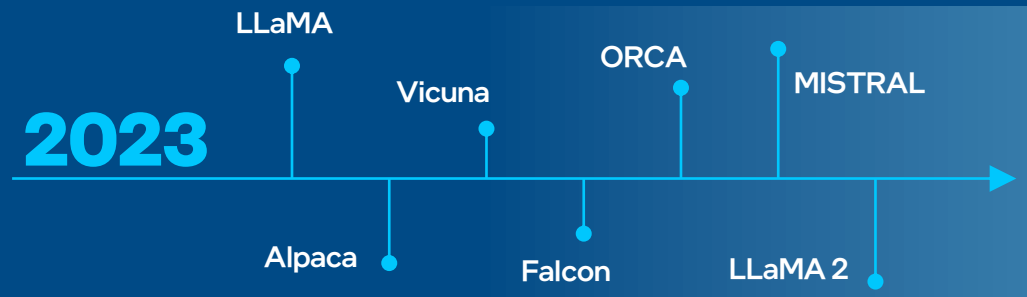
Studie von Epoch, University of Aberdeen, Center for the Governance of AI, University of St. Andrews, MIT, Eberhard Karls Universität Tübingen, Universidad Complutense

Es geht nicht nur um riesige Modelle

	Riesig (Drittanbieter)	VS.	**Klein und wendig** (10-100X)
Erklärbarkeit	Proprietäres Modell	VS.	Open-Source-basiertes Modell
Genauigkeit	All-in-One für allgemeine Zwecke	VS.	Zielgerichtet, domänenspezifisch, angepasst
Ort	Cloud-basiert (as-a-Service)	VS.	Lokale Ausführung von Inferenz; Edge, Client und lokal
Kosten	Kostenskalierung auf Dauer	VS.	Kostenmanagement
Schnelle Markteinführung	Schnelle Einrichtung (Sekunden)	VS.	Entwicklungszeit (Stunden/Tage)

Wachstum vieler kleinerer Modelle

100 Mrd. bis <20 Mrd. Parameter in 6 Monaten



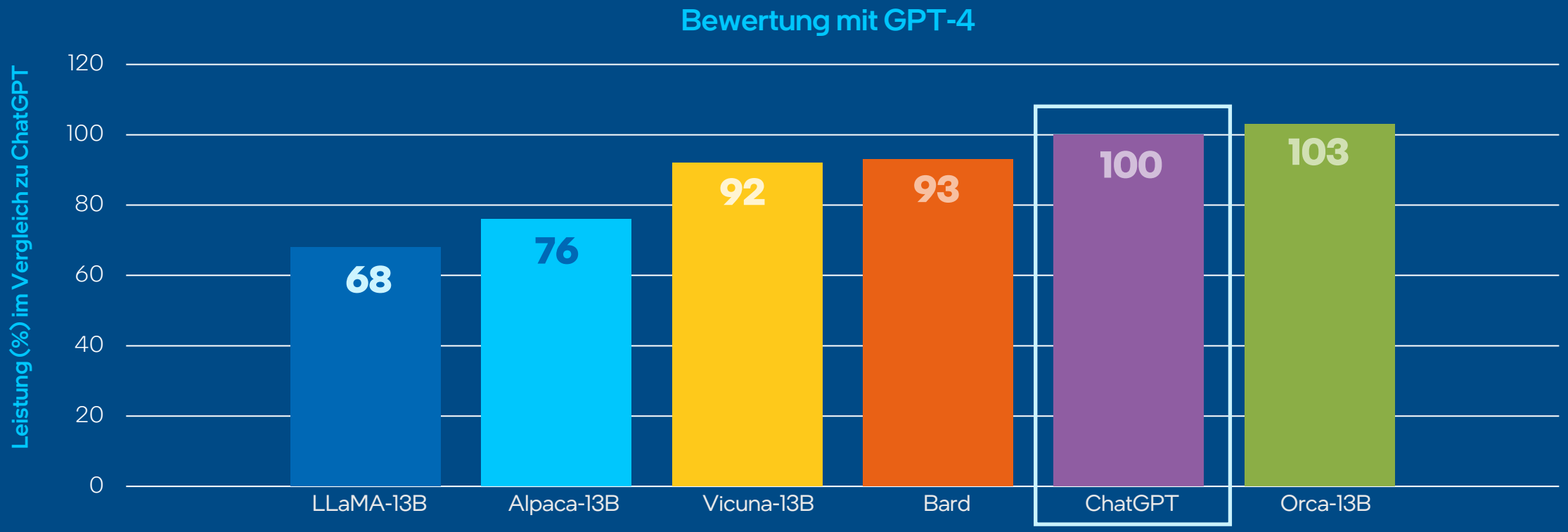
databricks

- Dutzende neue kleinere Modelle wöchentlich
- Kommerzielle und Open-Source-Lizenzen
- Hinweis darauf, dass kleinere Modelle die Genauigkeit größerer Modelle replizieren können, wenn sie mit sorgfältig gewonnenen Daten trainiert werden

- Tausende von domänenspezifischen kommerziellen Modellen und KI-Plattformen werden demonstriert
- Modelle können auf einigen wenigen Prozessoren in domänenspezifischen Daten fein abgestimmt werden

Kleinere Modelle funktionieren gut im Vergleich zu ChatGPT

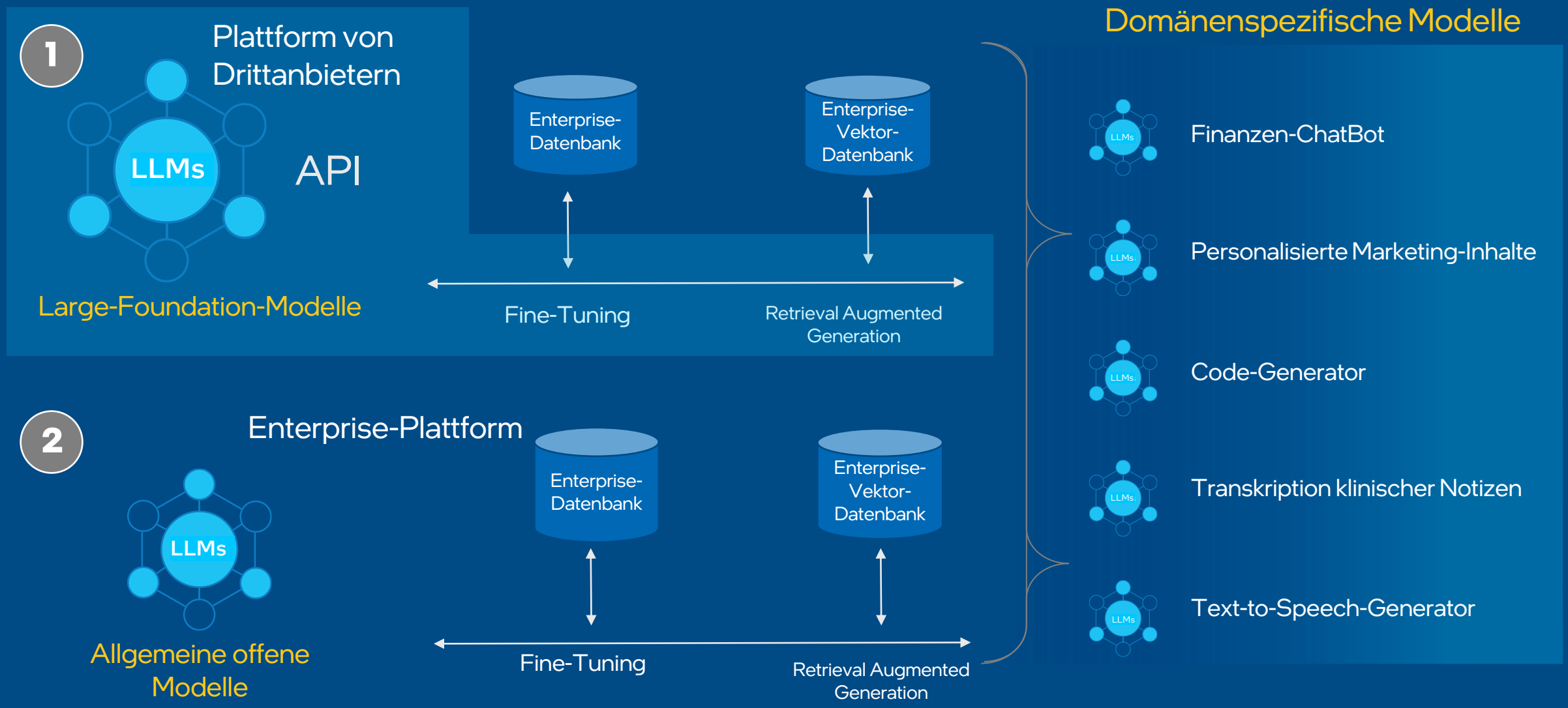
Beweis dafür, dass kleinere Modelle eine praktikable Option sind und im Vergleich zu großen Modellen wie ChatGPT immer noch gut funktionieren



Orca übertrifft eine Vielzahl von Grundlagen-Modellen, einschließlich OpenAI ChatGPT, wie von GPT-4 im Vicuna-Bewertungsset bewertet wurde.

Quelle: Microsoft Research (2023). Orca: Progressives Lernen aus komplexen Erklärungsspuren von GPT-4

Entwicklung domänenspezifischer Modelle



Domänenspezifische Modelle bieten viele Vorteile für Unternehmen

Kleinere, zielgerichtete Modelle können gleichwertige oder überlegene Leistung bieten und den ROI durch geringere Zeit- und Kosteninvestitionen erhöhen

Genauere Ausgabe

Verwenden Sie Ihre Unternehmensdaten für eine höhere domänenspezifische Genauigkeit

Kosteneinsparung

Feinabstimmung eines vortrainierten Modells und/oder Verwendung von RAG und Inferenzierung kleinerer Modelle

Überall auf der Plattform Ihrer Wahl bereitstellen

Lokale Ausführung von Inferenz; Edge, Client und lokal

Sicher und privat

Erfüllt Anforderungen an Datensicherheit und gesetzliche Vorschriften

Verantwortungsvolle KI

Möglichkeit des Modells, Datenquellen mit Feinabstimmung und RAG zu zitieren

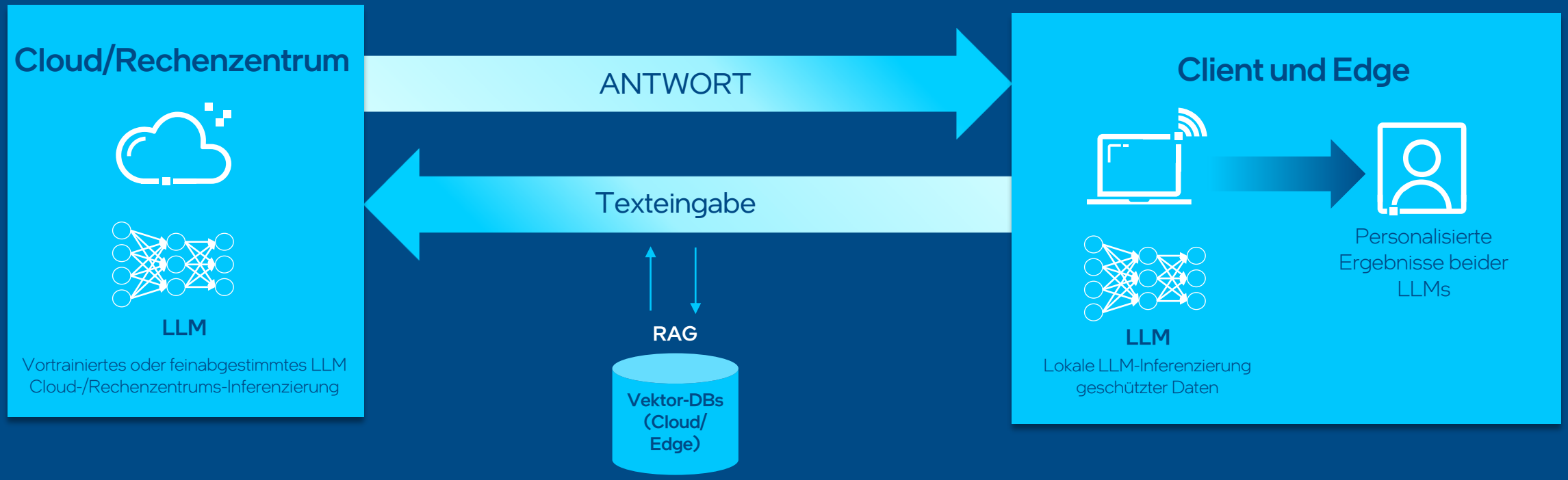
DIE ZUKUNFT

Es wird eine kleine Anzahl von riesigen Modellen und eine riesige Anzahl von kleinen, flexibleren KI-Modellen geben, die in unzählige Anwendungen eingebettet sind. [1]

[1]Quelle: Survival of the Fittest: Compact Generative AI Models Are the Future for Cost-Effective AI at Scale

Nahtlose Cloud-zu-Edge-KI-Plattform

Training und Inferenz in der Cloud. Höhere Domänen-Genauigkeit mit RAG.



Generative KI – ein Jahr in der Produktion

Die Verwendung domänenspezifischer, aber hochintelligenter Modelle nimmt zu

2022 EXPERIMENTIERUNG → **2023** PILOTPROGRAMME → **2024** PRODUKTION

Riesige Modelle ebneten den Weg

- Sehr effektiv für allgemeine Zwecke
- Training und Bereitstellung teuer
- Basierend auf großen öffentlichen Datenmengen
- Einfache Verwendung

Kleinere, domänenspezifische Modelle

- Verwenden Sie Ihre privaten Daten für geschäftsspezifische Ergebnisse
- Bereitstellung auf Ihrer Hardware
- Erhöhte Effizienz, Genauigkeit, Sicherheit und Rückverfolgbarkeit
- Zeit für die Entwicklung

Blog lesen

Survival of the Fittest: Compact Generative AI Models Are the Future for Cost-Effective AI at Scale

Intels Ansatz für domänenspezifische Modelle

Domänenspezifische Modelle

Vorteile

+ 10–100-mal kleinere Modelle bei gleichzeitiger Beibehaltung/Verbesserung der Genauigkeit
+ Wirtschaftlich bei Allzweck-Computing
+ Korrektheit, Quellenzuordnung, Erklärbarkeit
+ Verwendung privater/Unternehmensdaten
+ Kontinuierlich aktualisierte Informationen

Problemstellung

- Reduzierter Aufgabenbereich
- Erfordert Few-Shot-Feinabstimmung und Indexierung

INTEL ZIEL

Ermöglichung des kosteneffizientesten und allgegenwärtigsten Ansatzes für die Feinabstimmung und Bereitstellung 10.000er Modelle auf Intel Hardware unter Verwendung von Branchen-Frameworks, vortrainierten Modellen und Intel KI-SW und -Tools.

WEITERLESEN

Generative KI zum Greifen nah

E-Book ▪ Infografik

Enterprise KI: Hilfe bei der Überwindung von Einstiegshürden

Voraussetzungen	Wie eine Partnerschaft mit Intel helfen kann
Schnelle Markteinführung	Nutzen Sie die Entwickler-Ressourcen von Intel und Hugging Face, den Gaudi Developer Hub und die 5 Referenz-Kits, um einen Vorsprung in Sachen generativer KI zu erlangen.
Benutzererlebnis (Genauigkeit/Latenz)	Inferenz bei Modellen mit mehr als 10 Mrd. Parametern auf Intel® Gaudi® Beschleunigern und kleinen Modellen <20 Mrd. Parametern auf Intel® Xeon® Prozessoren mit Intel® AMX, die Benutzern ein Echtzeit-Erlebnis bieten.[1]
Rechenleistung-Verfügbarkeit	Intel® Xeon® CPU + Beschleuniger bieten eine kosteneffektive Alternative zu dem globalen GPU-Mangel. **Intel® Gaudi® 2 ist jetzt über SuperMicro verfügbar und bietet eine größere Verfügbarkeit für Intel® Gaudi® 3.**
Vertraute Technik	Inferenz kleinerer Modelle kann praktisch auf jeder Hardware durchgeführt werden, einschließlich allgegenwärtiger Lösungen, die bereits Teil Ihrer Recheneinstellungen sein könnten.[2]
Operationalisierung in großem Umfang	Intel® Gaudi® 2 bietet nahezu lineare Skalierbarkeit mit 24 100-GbE-Ports, die in jeden Beschleuniger integriert sind. Intel® Xeon® ist bereits in Ihrem Rechenzentrum, an fernen Standorten, in der Cloud bis zum Edge. **65 % der Rechenzentrum-Inferenzierung wird auf Intel® Xeon® ausgeführt.[3]**
Kosteneffektiv	In realen Arbeitsanwendungen stellt Intel die Branche auf den Kopf und demokratisiert KI, indem es bessere Leistung, niedrigere Preise und eine ausgewogenere Plattform für KI-Inferenz bietet. Artikel: NVIDIA zeigt, dass Intel® Gaudi eine 4-mal bessere Leistung pro Dollar bietet als H100.

[1]Quelle: Four Roadblocks to Implementing Generative AI

[2]Quelle: Survival of the Fittest: Compact Generative AI Models Are the Future for Cost-Effective AI at Scale

[3]Basierend auf einer Marktmodellierung von Intel für die weltweit installierte Basis von Rechenzentrumsservern, auf denen KI-Inferenz-Workloads laufen, Stand Dezember 2022.

Softwareressourcen für ein vereinfachtes Trainieren und Bereitstellen von generativer KI

Open-Source-Modell

- BLOOM (a BigScience initiative) — 176B (176B params, 59 languages, Open-access)
- BioGPT — Domäne 1.5B
- Stable Diffusion — Bild
- Llama2, GPT-J, MPT, Falcon — 7-65B LLM
- Stanford Alpaca — Fine-tuned 7B LLM
- Wissensdatenbank

Offene Software

- OpenVINO
- Intel® Extension für PyTorch (IPEX) — PyTorch
- Intel® Extension für Transformatoren (ITREX)
- Intel® Extension für DeepSpeed (IDEX) — DeepSpeed
- haystack
- fastRAG

GenAI-Plattform

- intel. Gaudi
- intel. Data Center GPU Max Series
- intel. Xeon
- intel Core Ultra

WEITERLESEN

Nutzen Sie generative KI für allgegenwärtige Hardware und offene Software

Wert maximieren

Vermeiden Sie die Herstellerbindung

Software mit Open-Source-Standards

MODIN · SciPy · dmlc XGBoost · PyTorch
pandas · NumPy · scikit learn · TensorFlow
Spark · Numba · SigOpt AutoML · oneAPI

Nutzen Sie das Hardware-Portfolio von Intel

optimiert für KI-Anwendungsfälle

intel Xeon · intel Core Ultra · intel Arc Graphics
intel Data Center GPU · intel Gaudi · intel Ethernet

Schaffen Sie neue Möglichkeiten vom Client und Edge bis hin zum Rechenzentrum und der Cloud mit Hardware, die durch Software und offene Standards für die KI von morgen optimiert wurde

Intel KI-Software-Portfolio

Ingenieurdaten → **Modelle erstellen** → **Optimieren und Bereitstellen**

MODIN · SciPy · pandas · NumPy · Spark · Numba

dmlc XGBoost · scikit learn · SigOpt AutoML

PyTorch · TensorFlow · intel Neural Compressor

ONNX RUNTIME · DirectML

OpenVINO · Write Once Deploy Anywhere

Datenanalysen bei Skalierung †

Maschinelle und Deep Learning-Frameworks, Optimierung und Bereitstellungstools†

oneAPI

- Intel® oneAPI Deep Neural Network Library
- Intel® oneAPI Collective Communications Library
- Intel® oneAPI Math Kernel Library
- Intel® oneAPI Data Analytics Library

Offenes, architekturübergreifendes Programmiermodell für CPUs, GPUs und andere Beschleuniger

CLOUD UND ENTERPRISE

CLIENT UND WORKSTATION

EDGE

intel AI ANALYTICS TOOLKIT

Beschleunigen Sie die End-to-End-Datenwissenschaft und KI

Intel® Tiber™ AI Cloud und Intel® Developer Catalog

Probieren Sie die neuesten Intel Tools und Hardware aus und greifen Sie auf optimierte KI-Modelle zu

Intel® Geti

Anmerkung/Schulung/ Optimierungsplattform

Hugging Face

Intel Optimierungen und Feinabstimmungsrezepte, optimierte Inferenzmodelle und Modellbereitstellung

Hinweis: Komponenten auf jeder Ebene des Stapels sind basierend auf erwarteten KI-Nutzungsmodellen für gezielte Komponenten auf anderen Ebenen optimiert, und nicht jede Komponente wird von den Lösungen in der rechten Spalte verwendet

† Diese Liste enthält beliebte Open-Source-Frameworks, die für Intel Hardware optimiert sind.

Vereinfachen Sie die Einführung von generativer KI in Unternehmen und reduzieren Sie die Zeit bis zur Produktion von ausgereiften, vertrauenswürdigen Lösungen

OPEA:

Vereinfachen Sie die Einführung von generativer KI in Unternehmen und reduzieren Sie die Zeit bis zur Produktion von ausgereiften, vertrauenswürdigen Lösungen

Open Platform for Enterprise AI

Partner von OPEA

anyscale

CLOUDERA

DATASTAX

kx

DOMINO

Hugging Face

Red Hat

drant

MariaDB

MINIO

sas

Yellowbrick

vmware

Wert von OPEA

- Hilft Unternehmen dabei, mit generativer KI (LLM, RAG) den Wert ihrer Daten schneller und einfacher zu nutzen
- Reduziert die Komplexität des fragmentierten Technologieumfelds und hilft bei der Skalierung von Lösungen in der Produktion
- Ermöglicht die Zusammenarbeit und Beiträge von Branchenführern, die mit der Linux Foundation kollaborieren

Effizient
Nutzt die vorhandene Infrastruktur, den KI-Beschleuniger oder andere Hardware Ihrer Wahl.

Perfekt abgestimmt
Lässt sich in Unternehmenssoftware mit heterogener Unterstützung und Stabilität im gesamten System und Netzwerk integrieren.

Offen
Bringt die besten Innovationen zusammen und ist frei von proprietärem Anbieter-Lock-in.

Allgegenwärtig
Läuft überall über eine flexible Architektur für Cloud, Rechenzentrum, Edge und PC.

Vertrauenswürdig
Bietet eine sichere Pipeline und Tools für Unternehmen, die für Verantwortung, Transparenz und Rückverfolgbarkeit sorgen.

Skalierbar
Bietet Zugriff auf ein lebendiges Partner-Ökosystem, um Ihre Lösung zu entwickeln und zu skalieren.

Hugging Face-Partnerschaft für generative KI

Hugging Face

Um das Trainieren sowie Innovationen im Bereich generativer und sprachlicher KI zu vereinfachen, hat sich Intel mit Hugging Face zusammengetan, einer beliebten Plattform für den Austausch von KI-Modellen und -Datensätzen. Hugging Face ist besonders bekannt für seine für NLP entwickelte Transformatoren-Bibliothek.

intel Xeon

Intel® hat mit Hugging Face an der Entwicklung modernster Hardware- und Softwarebeschleunigung zusammengearbeitet, um Transformatormodelle zu trainieren, fein abzustimmen und Vorhersagen damit zu treffen.

Die Hardwarebeschleunigung wird von skalierbaren Intel® Xeon® Prozessoren angetrieben, während die Softwarebeschleunigung auf unserem Portfolio an optimierten KI-Softwaretools, -Frameworks und -Bibliotheken beruht.

intel Gaudi

Intel® Gaudi® Deep-Learning-Beschleuniger werden über die Optimum Habana Library auch mit der Open-Source-Software Hugging Face kombiniert, um Entwicklern eine einfache Verwendung bei Tausenden von Modellen zu ermöglichen, die bereits von der Hugging Face-Community optimiert wurden.

Zudem hat Hugging Face verschiedene Bewertungen der Leistung von Intel® Gaudi® 2 bei generativen KI-Modellen veröffentlicht: Stable Diffusion, T5-3B, BLOOMZ 176B und 7B und das neue BridgeTower-Modell.

Intel®, Articul8 und BCG arbeiten zusammen, um sichere generative KI der Enterprise-Klasse zu bieten

Bahnbrechende Lösung basierend auf einem Intel KI-Supercomputer erschließt Geschäftswert mit benutzerdefinierten Datenmengen bei gleichzeitiger Wahrung eines hohen Maßes an Sicherheit und Datenschutz

Articul8* bietet eine schlüsselfertige GenAI-Software-Plattform, die Geschwindigkeit, Sicherheit und Kosteneffizienz bietet, um großen Unternehmenskunden bei der Operationalisierung und Skalierung von KI zu helfen. Die Plattform wurde auf der Grundlage von Intel® Hardware-Architekturen, einschließlich skalierbaren Intel® Xeon® Prozessoren und Intel® Gaudi® Beschleunigern, eingeführt und optimiert, wird aber auch eine Reihe von hybriden Infrastrukturalternativen unterstützen.

intel GAUDI

intel XEON

Im Anschluss an die frühe Bereitstellung der Technik bei der Boston Consulting Group (BCG) hat das Team die Plattform auf Unternehmenskunden in Industriesegmenten ausgeweitet, die ein hohes Maß an Sicherheit und domänenbezogenem Fachwissen erfordern, darunter Finanzdienstleistungen, Luft- und Raumfahrt, Halbleiter und Telekommunikation.

WEITERLESEN

Articul8-Ankündigung

Articul8-Website

Verantwortliche KI für Unternehmen

PROBLEMSTELLUNG:

Generative KI-Modelle lernen aus riesigen im Internet verfügbaren Datenmengen, die Vorurteile aus der Gesellschaft enthalten können – diese Vorurteile können versehentlich angewendet werden. LLMs können manipuliert werden, um Fehlinformation, Phishing-E-Mails oder Social-Engineering-Angriffe zu generieren oder zu verbreiten.

LLMs können oft „Halluzinationen“ haben und ungenaue Informationen generieren, was in Branchen wie dem Gesundheitswesen besonders problematisch sein kann, wo Modelle diagnostische und therapeutische Entscheidungen beeinflussen und Patienten potenziell schädigen können.

Weitere Informationen

Minimierung der Risiken generativer KI

LÖSUNGEN:

Unternehmen und Einzelpersonen, die mit KI-Technik arbeiten, müssen sicherstellen, dass ihre Software gemäß ethischen KI-Prinzipien entwickelt und bereitgestellt wird.

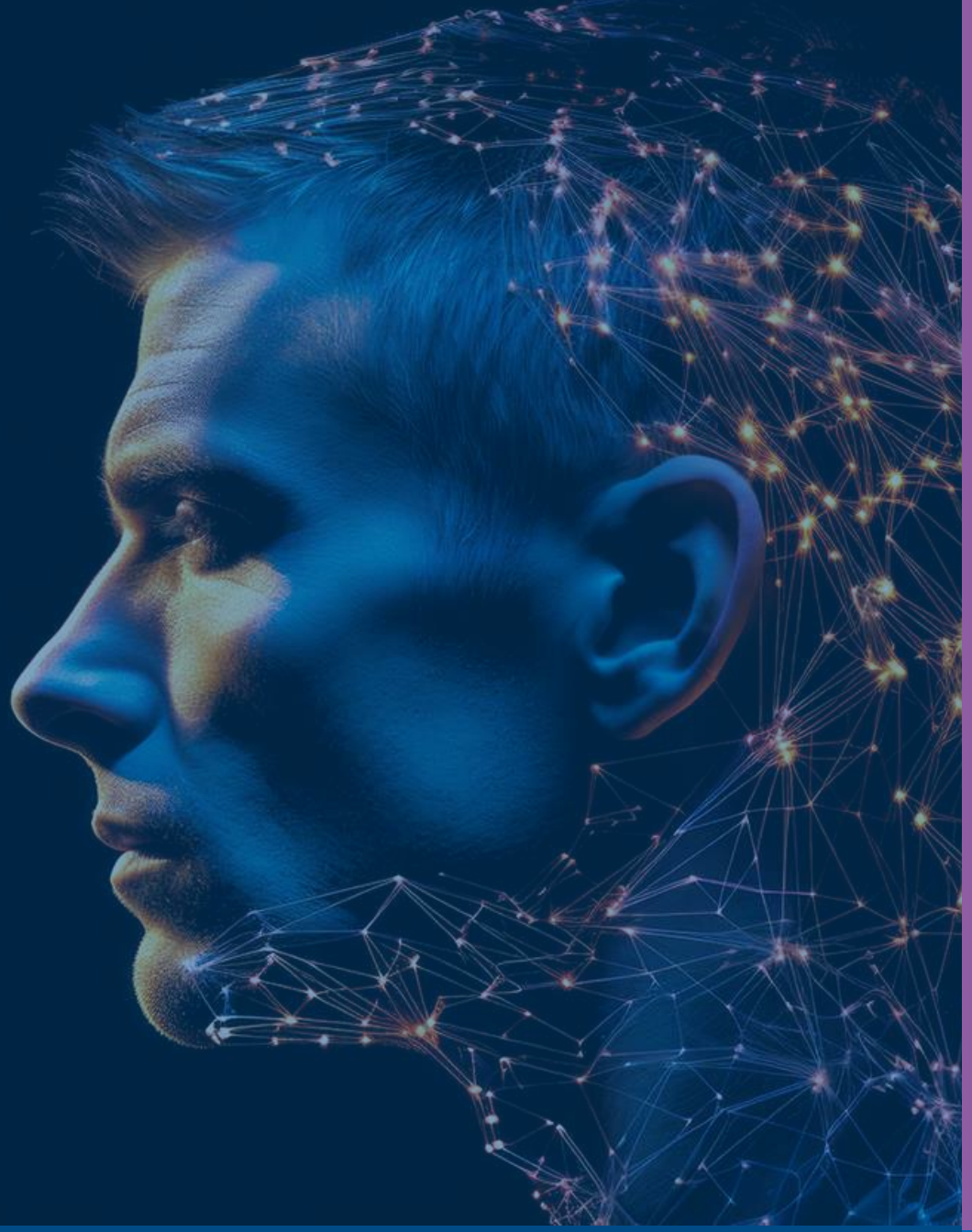
Die offenen Intel® Explainable AI Tools ermöglichen es Benutzern, Post-Hoc-Modelldestillation und -Visualisierung auszuführen, um das Vorhersageverhalten von TensorFlow*- und PyTorch*-Modellen zu untersuchen

LLMs werden in der Regel mit großen öffentlichen Datenmengen trainiert und dann mit potenziell vertraulichen Daten (z. B. Finanz- und Gesundheitswesen) feinabgestimmt.

Techniken wie Intel Open Federated Learning (OpenFL) integrieren Confidential Computing, sodass LLMs mit vertraulichen Daten sicher feinabgestimmt werden können, was wiederum die Generalisierbarkeit von Modellen verbessert und gleichzeitig Halluzinationen und Vorurteile reduziert.

Intel® Produkte für generative KI

KI überall verfügbar machen



Skalierbare Systeme für KI

Training und Feinabstimmung

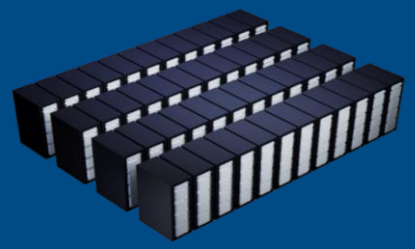
Schulung | Spitzen-Inferenz | Mainstream-Inferenz/Feinabstimmung | Ausgangs-Inferenz | Endpunkt-Inferenz

Inferenz und Bereitstellung

Cloud
Rechenzentrum

Edge

Client

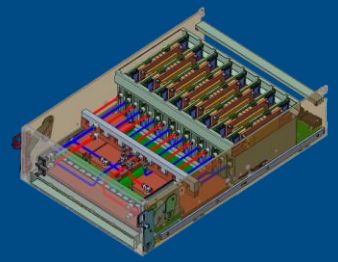


Cluster- und Rechenzentrumsskalierung

intel. GAUDI

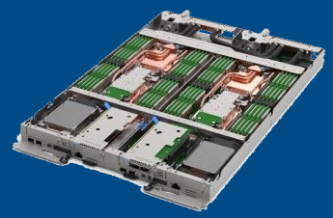
Multi-Knoten-Bereitstellung pro Rack

intel. GAUDI | intel. XEON | intel. XEON MAX SERIES



Multi-GPU oder Mehrprozessor-CPU

intel. XEON | intel. XEON MAX SERIES | intel. DATA CENTER GPU FLEX SERIES



Einzelprozessor-CPU oder einzelne GPU

intel. XEON | intel. DATA CENTER GPU FLEX SERIES | intel CORE ULTRA

Client-CPU

intel CORE ULTRA | intel. ARC GRAPHICS

intel. ETHERNET

Intel® Produkte für NLP/LLMs

Trainings-Inferenz

GAUDI®2

Intel® Gaudi® 2 KI-Beschleuniger wurden speziell für die Beschleunigung des Trainings und der Inferenz großskalierter Modelle wie LLMs und NLPs entwickelt.

Beschleunigung von generativer KI und Large Language Models mit Intel® Gaudi® 2

Intel® Gaudi® 2 bietet führende Leistung und optimale Kosteneinsparungen für KI-Training

Pressemitteilung

Der Intel® Gaudi® 2 Deep-Learning-Beschleuniger bietet wettbewerbsfähige Leistungen bei Deep-Learning-Training und -Inferenz, mit bis zu **2,4-mal schnellerer Leistung als NVIDIA A100**[1]

Newsroom ▪ Tech-Artikel

Jetzt ansehen

Intel® Webinar-Aufzeichnung, die die innovativen Funktionen des Intel® Gaudi® 2 KI-Prozessors bei der Erfassung des Potenzials von generativer KI und großen Sprachmodellen (LLMs) erörtert

Intel® Gaudi® 2 bleibt einzige Benchmark-Alternative zu NV H100 für GenAI-Leistung

[1]Die Leistungseigenschaften variieren je nach Verwendung, Konfiguration und anderen Faktoren. Workloads- und Konfigurationsdetails verfügbar unter: intel.com/performanceindex. Die Ergebnisse können von Fall zu Fall abweichen.

Gaudi2: Ideal für effizientes Training und Inferenz von Foundation-Modellen

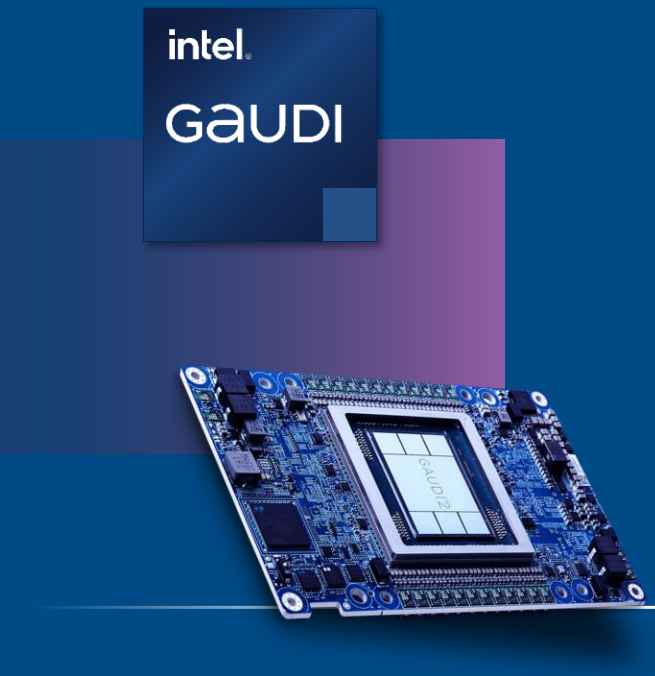
Gaudi2 ist für Deep-Learning-Leistung, Effizienz und Skalierbarkeit konzipiert, um die Anforderungen von großen Grundlagenmodellen wie LLMs (GPT) und GAIs (Stable Diffusion) zu erfüllen

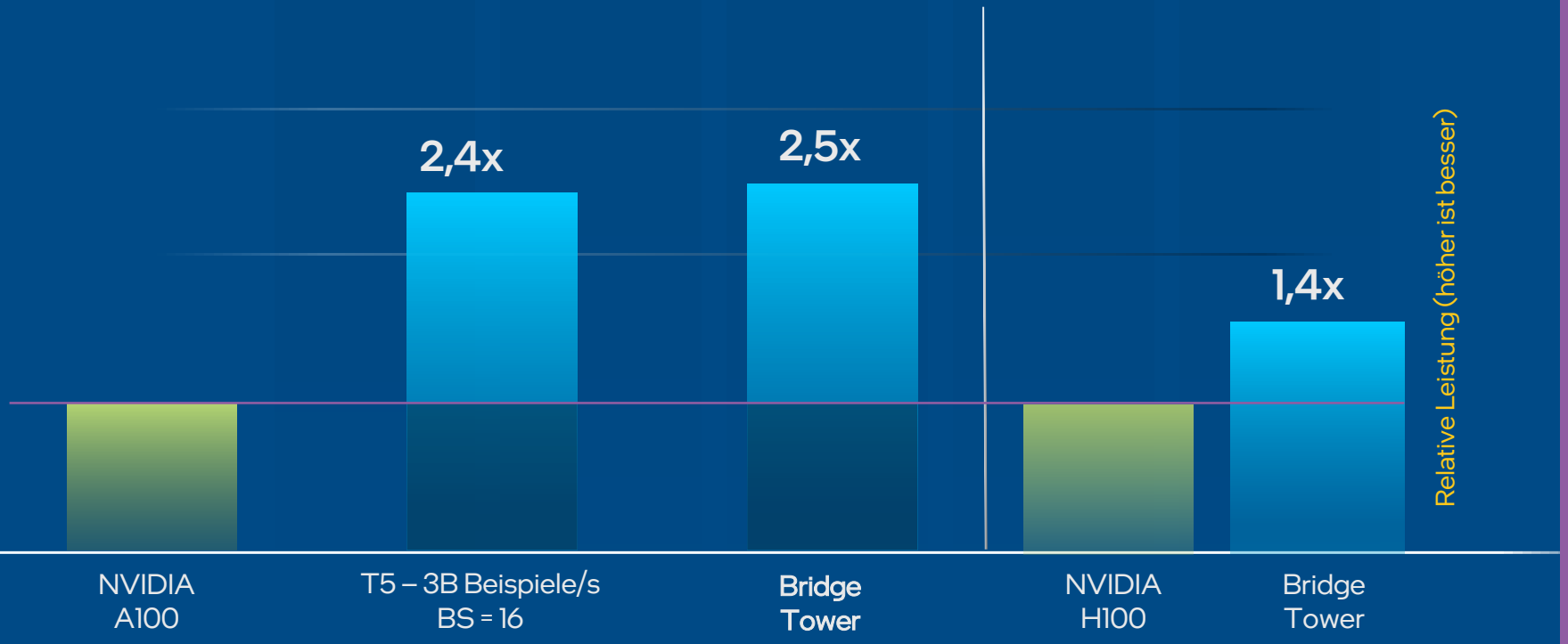
Voraussetzungen	Gaudi2
Frequenz	**1,5-2-mal schneller** als A100 für Training und Inferenz
Arbeitsspeicher	Jedes Gaudi2-Gerät verfügt über **96 GB On-Chip-Arbeitsspeicher mit hoher Bandbreite**, was es einfacher macht, große Grundlagenmodelle im Arbeitsspeicher zu integrieren und sie in großem Umfang zu trainieren und bereitzustellen
Skalierbarkeit	Skalierungseffizienz mit **24 x 100 GbE-Ports auf Chip integriert**, direkte All-to-All-Konnektivität zwischen 8 Karten in einem Server und offene ROCEv2-basierte Kommunikation innerhalb und über Server hinweg.
Benutzerfreundlichkeit	Migrieren oder entwickeln Sie Modelle mit **minimalen Code-Änderungen** mit SynapseAI, PyTorch und DeepSpeed
Energieeffizienz	**~1,8-mal höherer Durchsatz/Watt im Vergleich zu A100**
Kosteneffizienz	Basierend auf der speziell entwickelten Gaudi Architektur der 1. Generation, die bis zu **40 % bessere Preisleistung** als A100 in der Amazon-Cloud liefert

Quelle: https://habana.ai/habana-claims-validation/

Feinabstimmung für zahlreiche LLMs

Hugging-Face-Evaluationen belegen die LLM-Leistung des Intel® Gaudi® 2 Beschleunigers im Vergleich zu NVIDIA A100 und H100





Besuchen Sie https://habana.ai/habana-claims-validation für Workloads und Konfigurationen. Die Ergebnisse können von Fall zu Fall abweichen.

https://huggingface.co/blog/habana-gaudi-2-benchmark
https://huggingface.co/blog/bridgetower

GPT-J: Intel® Gaudi® 2 Ergebnisse

Die Ergebnisse der Intel® Gaudi® 2 Inferenzleistung für GPT-J sind eine überzeugende Bestätigung für seine hervorragenden Leistung.

- Die Intel® Gaudi® 2 **Inferenzleistung** auf GPT-J-99 und GPT-J-99.9 für Serverabfragen und Offline-Stichproben beträgt **78,58 pro Sekunde bzw. 84,08 pro Sekunde.**[1]
- **Intel® Gaudi® 2 bietet eine überzeugende Leistung im Vergleich zu NVIDIA H100**, wobei H100 einen leichten Vorteil von 1,09-fache (Server)- und 1,28-fache (Offline)-Leistung im Vergleich zu Gaudi 2 aufweist.[1]
- **Intel® Gaudi® 2 übertrifft NVIDIA A100 um das 2,4-Fache (Server) und das 2-Fache (Offline).**[1]
- Die Intel® Gaudi® 2 Einreichung verwendete FP8 und erreichte **eine Genauigkeit von 99,9 %** bei diesem neuen Datentyp.[1]

WEITERLESEN

Intel® Gaudi® 2 Software-Updates werden alle sechs bis acht Wochen veröffentlicht, sodass Intel auch weiterhin Leistungsverbesserungen und eine erweiterte Modellabdeckung bei MLPerf-Benchmarks erwartet.

Newsroom-Artikel

MLCommons-Ankündigung

[1]Die Leistung variiert je nach Verwendung, Konfiguration und anderen Faktoren. Workloads und Konfigurationsdetails verfügbar unter: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ Die Ergebnisse können von Fall zu Fall abweichen.

Intel® Gaudi® 2: Benchmark-Ergebnisse

Benchmark-Ergebnisse von Supermicro, der branchenweit erste Intel® Gaudi® 2 OEM

Gaudis Validierung von Ansprüchen

LLM-Training und -Inferenz mit Intel® Gaudi® 2 KI-Beschleunigern

Benchmarks

Schnelleres Training und Inferenz: Intel® Gaudi® 2 im Vergleich zu NVIDIA A100 80 GB

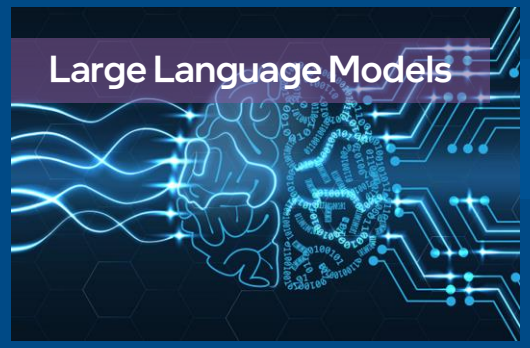
Benchmarks

Die Ergebnisse können von Fall zu Fall abweichen.

Intel® Gaudi® 2: Grundlegendes Modell-Training und Inferenz

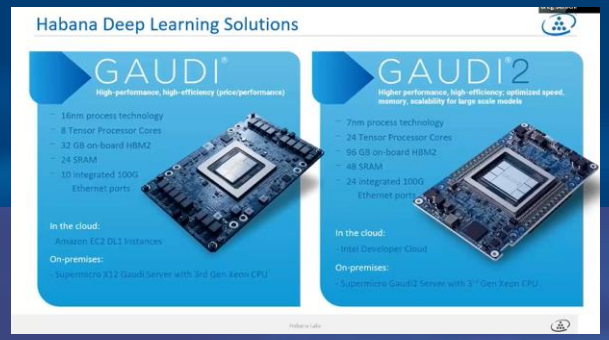
Verfügbare Gaudi-taugliche Modelle gibt es im Entwickler-Katalog

GAUDI®2



Intel® Gaudi® Schulung für Entwickler

[Erste Schritte: Deep Learning und Inferenz mit Gaudi]



[Maximierung der Leistung von Intel® Gaudi® 2: Beschleunigung von generativer KI und Large Language Models]

[Maximierung der Modellleistung mit Intel® Gaudi® Prozessoren: Erweiterte Tools und Strategien für optimale Ergebnisse]

Intel® Gaudi® Software (SynapseAI® Software Suite)

Vereinfachte Entwicklung: So geht es am besten

Ziel: Vereinfachung der Migration vorhandener Software auf Intel® Gaudi® KI-Beschleuniger, Erhalt von Softwareinvestitionen und einfache Entwicklung neuer Modelle – für das Training und die Bereitstellung der zahlreichen und wachsenden Modelle, die Deep Learning, generative KI und Large Language Models definieren.

Umfangreiche Unterstützung für Datenwissenschaftler, Entwickler, und IT- und Systemadministratoren mit:

- Entwickler-Website
- GitHub

Intel® Gaudi® KI-Beschleuniger

Das Software-Technologieumfeld für Deep Learning bringt führende Softwareanbieter, Tools und Code zusammen, um die Entwicklung modernster Deep-Learning-Modelle basierend auf den Frameworks PyTorch, TensorFlow, PyTorch Lightning und DeepSpeed zu beschleunigen.

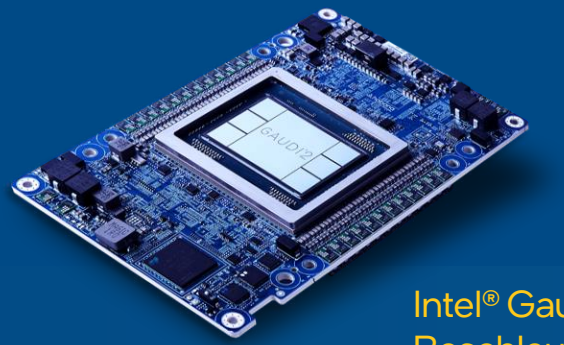
cnvrg.io

PyTorch Lightning

Bereit die Intel® Gaudi® Software zu verwenden?

Intel® Gaudi® 2 KI-Beschleuniger JETZT VERFÜGBAR! Exklusiv in der Denvr Cloud

Intel® Gaudi® 2 Software-Ökosystem



Intel® Gaudi® 2 KI-Beschleuniger (7 nm)

Intel® Gaudi® 2 – Ideal für die Anforderungen von generativer KI

- Jetzt erhältlich! Gaudi 2 Cluster in der Denvr Cloud
- Bis zu 8 Gaudi 2 Knoten testen
- Priority VIP-Preise für Intel Kunden
- Denvr Dataworks High-Touch Commercial Service und Support
- Nahtlose Migration auf Gaudi 2 Cluster in der Denvr Cloud
- Exklusive Prioritätsposition für Gaudi 3 Cluster in der Denvr Cloud – bald verfügbar!

JETZT LOSLEGEN

DEMNÄCHST

intel GAUDI

Trainings-Inferenz

Intel® Gaudi® 3

Die Leistung, Skalierbarkeit und Effizienz von Intel® Gaudi® 3 Beschleunigern ermöglichen mehr Kunden eine größere Auswahl und verhelfen Unternehmen zu neuen Erkenntnissen, Innovationen und Umsätzen.

DEMNÄCHST – Intel® Gaudi® 3 KI-Beschleuniger

Mehr Auswahl für GenAI mit Leistung, Skalierbarkeit und Effizienz

Intel® Gaudi® 3 wird einen bedeutenden Sprung hinsichtlich KI-Training und Inferenz für globale Unternehmen bieten, die GenAI in großem Umfang bereitstellen möchten.

Pressemitteilung

Intel® Gaudi® 3 Beschleunigerleistung im Vergleich zu NVIDIA H100

Intel® Gaudi® 3 wird voraussichtlich

eine um durchschnittlich 50 % schnellere Trainingszeit[3] bei Llama2-Modellen mit 7B und 13B Parametern und GPT-3 175B Parametermodell bieten.

Intel® Gaudi® 3 wird H100 voraussichtlich übertreffen:

50 % besserer Beschleuniger-Inferenz-Durchsatz[1]

40 % bessere Inferenz-Energieeffizienz[2]

für Llama 7B und 70B Parameter und Falcon 180B Parametermodelle

WEITERLESEN

WHITEPAPER

Intel® Gaudi® 3 wird ab dem 2. Quartal 2024 für OEMs verfügbar sein, darunter:

Lenovo DELL Hewlett Packard Enterprise SUPERMICRO

[1]NV H100 Vergleich basierend auf https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8. Die angegebenen Zahlen sind pro GPU. Gegenüber Intel® Gaudi® 3 Prognosen für LLAMA2-7B, LLAMA2-70B und Falcon 180B. Die Ergebnisse können von Fall zu Fall abweichen.
[2]NV H100 Vergleich basierend auf https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8. Die angegebenen Zahlen sind pro GPU. Gegenüber Intel® Gaudi® 3 Prognosen für LLAMA2-7B, LLAMA2-70B und Falcon 180B. Energieeffizienz für NVIDIA und Gaudi 3 basierend auf internen Schätzungen. Die Ergebnisse können von Fall zu Fall abweichen.
[3]NV H100 Vergleich basierend auf: https://developer.nvidia.com/deep-learning-performance-training-inference/training, Reiter „Large Language Model" im Vergleich zu Intel® Gaudi® 3 Prognosen für LLAMA2-7B, LLAMA2-13B und GPT3-175B vom 28.3.2024. Die Ergebnisse können von Fall zu Fall abweichen.

Intel® Produkte für NLP/LLMs

Inferenz

Skalierbare Intel® Xeon® Prozessoren der 4. und 5. Generation beschleunigen NLP mit Intel® DL Boost, Intel® AMX und Intel® AVX-512. Sie sind für High-Performance-Computing konzipiert und können zur Beschleunigung von NLP-Workloads verwendet werden. Sie können eine große Anzahl von Threads, eine große Speicherkapazität und eine hohe Speicherbandbreite bewältigen, was sich für NLP-Workloads wie Sprachübersetzung, Textzusammenfassung und Text-to-Speech eignet.

Intel® Xeon® der 5. Generation: Der für KI entwickelte Prozessor

Mit KI-Beschleunigung in jedem Kern erfüllen Intel® Xeon® Prozessoren der 5. Generation anspruchsvolle End-to-End-KI-Workloads, bevor Kunden separate Beschleuniger hinzufügen müssen

Höhere Leistung bei KI-Inferenz

bis zu

42 %

im Vergleich zur vorherigen Generation[1]

Allgemeine Rechenleistungsgewinne

durchschnittlich

21 %

im Vergleich zur vorherigen Generation[1]

Schnellere natürliche Sprachverarbeitung

bis zu

23 %

im Vergleich zur vorherigen Generation[1]

Sandra Rivera, Intel Executive Vice President und General Manager der Data Center and AI Group

„Unsere Intel® Xeon® Prozessoren der 5. Generation wurden für KI entwickelt und bieten Kunden mehr Leistung, die KI-Funktionen in Cloud-, Netzwerk- und Edge-Anwendungsfällen bereitstellen. Als Ergebnis unserer langjährigen Arbeit mit Kunden, Partnern und dem Entwickler-Ökosystem starten wir Intel® Xeon® der 5. Generation auf einer bewährten Grundlage, die eine schnelle Einführung und Skalierung bei niedrigeren Gesamtbetriebskosten ermöglichen wird."

Weitere Informationen

Website

Produktbeschreibung

Für Workloads und Konfigurationen, besuchen Sie intel.com/processorclaims: Skalierbare Intel® Xeon® Prozessoren der 5. Generation. Die Ergebnisse können von Fall zu Fall abweichen.

Intel® Xeon®: CPU-Leistungsführerschaft in KI-Anwendungen der realen Welt

In realen Arbeitsanwendungen revolutioniert Intel die Branche und demokratisiert KI, indem es eine bessere Leistung, einen niedrigeren Preis und eine ausgewogenere Plattform für KI-Inferenz bietet mit:

- Größerer Cache, der mit Datenlokalität und großer Speicherkapazität hilft und größere Probleme lösen kann
- Höhere Kernfrequenz, mehrere Skalar-Ports und Out-of-Order-Ausführung, die zur Beschleunigung von Single-Threaded- oder Multi-Threaded-, aber skalaren Computing-Prozessen beitragen
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512), das bei der Nicht-DL-Vektorberechnung hilft
- Intel® Advanced Matrix Extensions (Intel® AMX), das integrierte Hardware-Unterstützung
  für KI-Beschleunigung ist

Vollständiger Tech-Artikel

Infografik

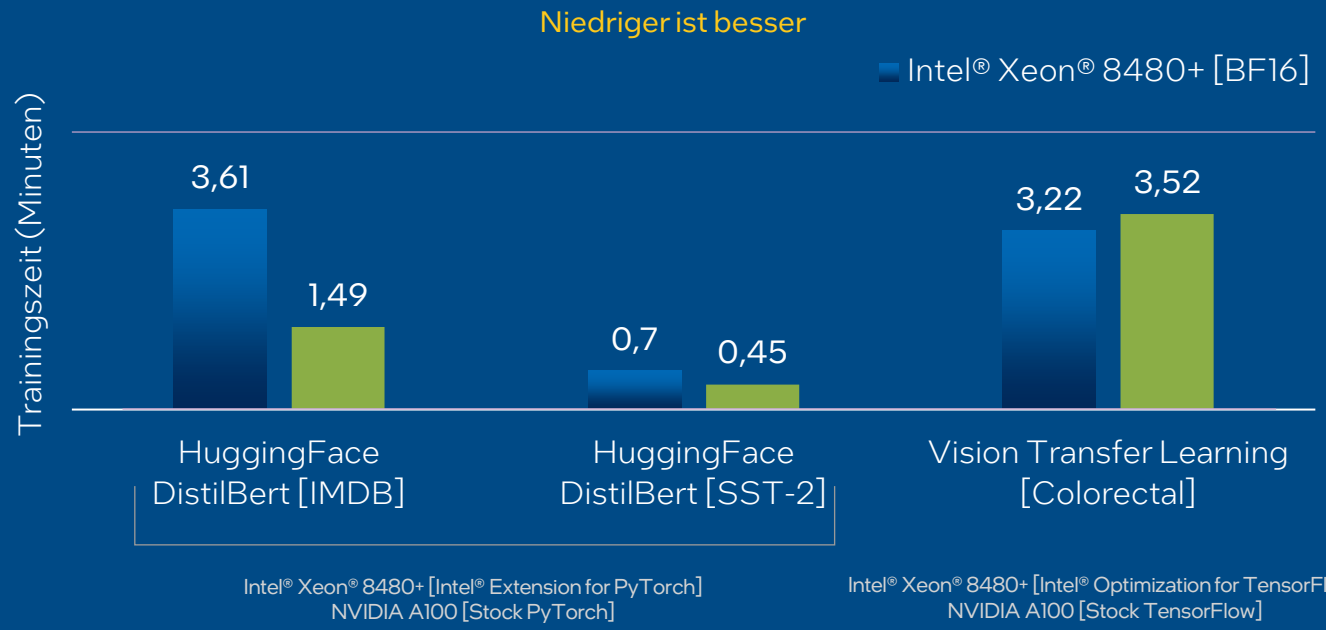
intel XEON

Entwirrung des GPU-Mythos: Wie CPUs mit integrierten Beschleunigern KI revolutionieren

Feinabstimmung von Modellen in weniger als 4 Minuten mit skalierbaren Intel® Xeon® Prozessoren[1]

Hugging Face

Feinabstimmung der Trainingszeit-Leistung von Intel® Xeon® Platinum 8480+ Prozessor im Vergleich zu NVIDIA A100 GPU



Siehe auch:

Bessere Leistung: Numenta auf Intel® CPUs im Vergleich zu NVIDIA GPUs

[1]Siehe [A221] des Leistungsindex für skalierbare Intel Xeon Prozessoren der 4. Generation Die Ergebnisse können von Fall zu Fall abweichen.

LLMs auf Intel® Xeon® Prozessoren der 4. Generation

Künstliche Intelligenz (KI)-Chatbot-Technik wird bei Unternehmen und Organisationen immer beliebter, um mit Kunden zu interagieren und den Kundendienst zu verbessern. Die Entwicklung, Optimierung und Wartung von Chatbots für bestimmte Anwendungsfälle ist jedoch teuer und kann für viele Unternehmen unerschwinglich sein.

WEITERE INFORMATIONEN

Tuning-Leitfaden für KI auf skalierbaren Intel® Xeon® Prozessoren der 4. Generation

Link zum Leitfaden ›

Intel® Xeon® Prozessoren der 4. Generation bieten verbesserte Datenverwaltung und effiziente Berechnungen durch die **Intel® Advanced Matrix Extensions (AMX)**. In Kombination mit der über die Intel® Extension für PyTorch verfügbaren **Auto Mixed Precision (AMP)**-Funktionalität wird dieser Technik-Stack für Workloads wie Transfer Learning und das Training kleiner/mittlerer Modelle von Grund auf wettbewerbsfähig.

Anleitung – Tech-Artikel

Cisco UCS mit Intel® Xeon® Prozessoren der 5. und 4. Generation für generative KI

Kleiner ist besser: Q8-Chat LLM ist ein effizientes generatives KI-Erlebnis auf Intel® Xeon® Prozessoren

LLMs erfordern viel Rechenleistung, die in der Regel in High-End-GPUs zu finden ist, um schnell genug für Anwendungsfälle mit geringer Latenz wie Such- oder Gesprächsanwendungen Vorhersagen zu treffen. Leider sind die damit verbundenen Kosten für viele Unternehmen unerschwinglich und erschweren den Einsatz moderner LLMs in ihren Anwendungen.

Hugging Face

„Mehr Unternehmen wären besser bedient, sich auf kleinere, bestimmte Modelle zu konzentrieren, dessen Training und Ausführung billiger sind."

Erfahren Sie mehr über Optimierungstechniken, die die LLM-Größe und Inferenz-Latenz reduzieren und es ermöglichen, dass sie effizient auf Intel® CPUs ausgeführt werden.

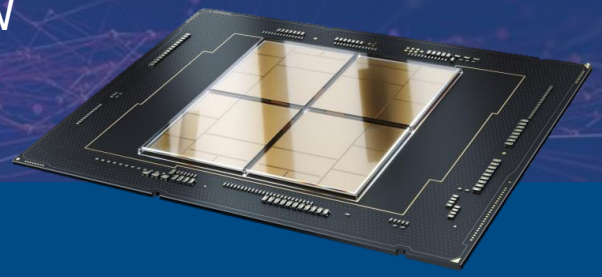
Anleitung – Tech-Artikel ›

Jetzt loslegen mit Intel® Xeon® der 4. Generation und Hugging Face

Intel® Xeon® Prozessoren für LLMs

ZUSAMMENFASSUNG

- Gut positioniert für Inferenzierung spezieller Domänen-LLMs
- Hervorragend für Transfer-Learning-Anwendungsfälle
- Stellen Sie LLMs auf Intel® Xeon® mit Open-Source-SW bereit, um optimale Leistung zu erhalten



Skalierbare Intel® Xeon® Prozessoren für LLMs

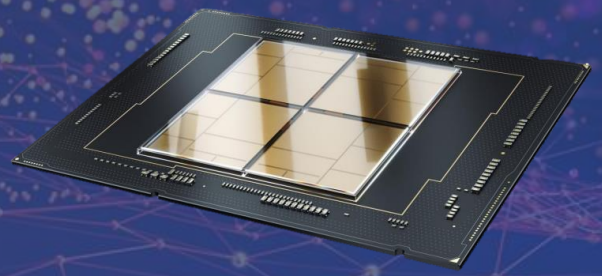
Ideal für die Entwicklung und Bereitstellung von Allzweck-KI-Workloads mit den beliebtesten KI-Frameworks und -Bibliotheken

- Nutzen Sie vorhandene Infrastruktur für Inferenzierung domänenspezifischer LLMs
- Hervorragend für Transfer-Learning-Anwendungsfälle
- Stellen Sie LLMs auf Intel® Xeon® mit Open-Source-SW bereit, um optimale Leistung zu erhalten

Intel® Xeon®

Führende CPU-Leistung bei realen KI-Anwendungen

Tech-Artikel ▪ Infografik



GPT-J

Intel® Xeon® der 4. Generation – Ergebnisse

2 Absätze pro Sekunde im Offline-Modus[1]

1 Absatz pro Sekunde im Echtzeit-Server-Modus[1]

Newsroom-Artikel ▪ MLCommons-Ankündigung

Entwirrung des GPU-Mythos: Wie CPUs mit integrierten Beschleunigern KI revolutionieren

Alibaba NLP Fallstudie zu Intel® Xeon® der 4. Generation mit Intel® AMX

WEITERLESEN

[1]Die Leistung variiert je nach Verwendung, Konfiguration und anderen Faktoren. Workloads und Konfigurationsdetails verfügbar unter: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/. Die Ergebnisse können von Fall zu Fall abweichen.

Intel® Produkte für NLP/LLMs

Inferenz im kleinen Maßstab auf dem Client

Intel® Core™ Ultra leitet das Zeitalter des KI-PCs ein

Intel® Core™ Ultra Prozessoren sind für erstklassige flache und leistungsstarke Laptops optimiert und bieten 3D-Performance-Hybridarchitektur, fortschrittliche KI-Funktionen und sind mit integrierter Intel® Arc™ GPU verfügbar. Die mit dem neuen Intel® 4 Prozess entwickelten Intel® Core™ Ultra Prozessoren bieten ein optimales Verhältnis zwischen Leistung und Energieeffizienz für Gaming, Content-Gestaltung und Produktivität unterwegs.

Anwendungsfälle: KI auf dem PC

Content-Gestaltung: Foto- und Videosuche und -bearbeitung

Schnellere natürlichere Filter, qualitativ hochwertigere Vorschauen und schnellere Exportzeiten mit automatisierten, schnelleren Suchen.

Unterhaltsame PC-Spiele

Neue KI-Funktionen für In-Game, 3D-Animation für zusätzlichen Realismus, Transkription und Chat-Übersetzung.

Content-Gestaltung: Text zu Bild

Neue KI-Effekte und Funktionen für die Erstellung von Bildern mit nur wenigen beschreibenden Wörtern – Marketing, Werbung, Design.

KI auf dem PC

„Das Banale erschließen“

Zusammenarbeit/Streaming

Neue KI-Funktionen für Videokonferenzen, Streaming und Zusammenarbeit der nächsten Generation, die die Akkulaufzeit bewahren.

Produktivität

KI-Assistenten für das Schreiben, Erstellen, Kodieren und Offline-Funktionen wie Text- und Grammatikvorhersage.

Zugänglichkeit

KI-gestützte audiovisuelle Funktionen für verschiedene Benutzeranforderungen, was es einfacher macht, auf dem PC zu erstellen und produktiv zu sein.

*Intel 14. Dezember 2023, Pressemitteilung zu AI Everywhere Event

Intel® Core™ Ultra für generative KI

Der energieeffizienteste Client-Prozessor von Intel läutet das Zeitalter des KI-PCs ein

Erhebliche Verbesserungen bei der Effizienz und Leistung

KI-EFFIZIENZ

bis zu **70 %**

schnellere generative KI-Leistung[2]

ENERGIEEINSPARUNG

bis zu **25 %**

Reduzierung des Energieverbrauchs[3]

WEITERLESEN Ankündigung ▪ Produktbeschreibung ▪ Website

Intel® Core™ Ultra verfügt über den ersten Client-On-Chip-KI-Beschleuniger von Intel – die Neural Processing Unit oder NPU –, um ein neues Niveau an energieeffizienter KI-Beschleunigung mit **einer 2,5-mal besseren Energieeffizienz** als die vorherige Generation zu ermöglichen[1]

Die Intel® Core™ Ultra Generationen H und U umfassen beide zwei neue Low Power Island (LP-E)-cores für Workloads mit geringer Intensität mit zwei neuralen Computing-Engines innerhalb der Intel KI-NPU, die für **generative KI-Inferenzierung** entwickelt wurden.

[1]Gemessen anhand Leistung/Watt mit UL Procyon KI-Benchmark bei der Ausführung eines int8-Modells mit Intel® Core™ Ultra 7 165H NPU im Vergleich zu Intel® Core™ i7-1370P GPU.
[1,2,3]Workloads und Konfigurationen unter www.intel.com/performanceindex. Die Ergebnisse können von Fall zu Fall abweichen.

Beschleunigung von KI-Innovationen

Intel® arbeitet mit führenden Branchen-ISVs zusammen, um Ihr Erlebnis mit der KI zu optimieren.

Das AI PC Acceleration Program soll unabhängige Hardware-Anbieter (IHVs) und unabhängige Softwarehersteller (ISVs) mit Intel® Ressourcen zusammenbringen, einschließlich KI-Toolchains, Schulungen, Co-Engineering, Software-Optimierung, Hardware, Design-Ressourcen, technischem Fachwissen, Co-Marketing und Vertriebsmöglichkeiten.

Weitere Informationen

Beschleunigen Sie die KI-Entwicklung in Unternehmen mit der Intel® Tiber™ AI Cloud

(ehemals Intel® Developer Cloud)

Erstellen Sie Prototypen und trainieren, testen und führen Sie Anwendungen und Workloads auf einem Cluster mit der neuesten Intel® Hardware und Software aus.

Beschleunigen und **skalieren Sie KI** mit den neuesten Hardware- und Softwareinnovationen in dieser Entwicklungsumgebung. **Profitieren Sie von mehr Rechenleistung** und Auswahlmöglichkeiten zur **Feinabstimmung Ihrer Software** und **generativen KI.**

Jetzt loslegen mit Intel

Machen Sie praktische Erfahrungen mit den neuesten Intel Produkten. Stärken Sie Ihre KI-Fähigkeiten mit Intel.

Frühzeitiger Technikzugriff

Bewerten Sie Vorabversionen der Intel Plattformen und der zugehörigen für Intel Technologie optimierten Software-Stacks.

KI in großem Maßstab bereitstellen

Beschleunigen Sie KI-Bereitstellungen mit den neuesten Toolkits für maschinelles Lernen von Intel und in der Intel® Tiber™ Developer Cloud gehosteten Bibliotheken.

Technischen Artikel lesen ›

Jetzt loslegen ›

Handlungsaufforderung

BILDUNGSBEREICH

Verstehen Sie, wie Intel® Technik für generative KI und domänenspezifische Modelle eingesetzt werden kann und in welchem Umfang die Intel® Xeon® und Intel® Gaudi® Produktlinien Ihnen helfen können, neue Marktchancen zu erschließen.

Einstieg

ENGAGEMENT

Legen Sie los mit

Intel® Tiber™ AI Cloud

Beschleunigen und skalieren Sie KI mit den neuesten Hardware- und Softwareinnovationen in dieser Entwicklungsumgebung

und

Nutzen Sie KI-Referenzkits

KONTAKT

Weitere Informationen erhalten Sie bei Ihrem **Intel® Vertreter.**

Zugriff auf den Intel® Partner Alliance Kundensupport

Intel® Virtual Assistant

Dieser Chat-Bot befindet sich in der unteren rechten Ecke jeder Partner Alliance-Webseite und bietet Selbsthilfe bei den meisten Fragen oder einen schnellen Link zu einem Live-Support-Mitarbeiter.

Blade „Hilfe erhalten“

Senden Sie eine Online-Support-Anfrage.

Dieser Link befindet sich in der Fußzeile der meisten Seiten auf dem Partner Alliance-Portal.

Partner Alliance-Seite „Hilfe erhalten“

Die Seite Hilfe erhalten bietet detaillierte Selbsthilfe-Leitfäden zu den meisten Tools und Leistungen, die Mitgliedern der Partner Alliance zur Verfügung stehen.

KI-Aktivierungszonen

Digital-First-KI-Workspaces, die wichtige Ressourcen, Tools und Vorteile vereinen und Partner dazu anregen, Lösungen auf der Grundlage von Intel® Technik zu entwickeln, zu vermarkten und zu verkaufen

Technische Möglichkeiten

Verkaufs- und Marketing-Enablement

Technische Möglichkeiten

Verkaufs- und Marketing-Enablement

Technische Möglichkeiten

Verkaufs- und Marketing-Enablement

KI-Referenz-Kits

Der Einsatz dieser Referenzkits ermöglicht es Unternehmen, die Zeit bis zur Lösungsfindung **deutlich zu verkürzen** und erhebliche **Leistungssteigerungen zu erzielen.**

Finanz- und Versicherungs-wesen

Betrugsaufdeckung

GitHub ▪ Blog ▪ Blueprint

Gesundheitsbereich und Biowissenschaften

Krankheitsschutz

GitHub ▪ Blog

Fertigung und Versorgungs-unternehmen

Erkennen von Anomalien

GitHub ▪ Blog

Flotten-management

Prädiktive Wartung

GitHub

Prozessauto-matisierung

Dokumentenautomatisierung

GitHub ▪ Blog ▪ Blueprint

Workflows

- DL Transfer Learning
- HF-Feinabstimmung und Inferenz-Optimierung
- DL-verteilte Komprimierung
- Verteilter klassischer ML-Workflow
- DL-Vorab-Training mit Intel® Beschleunigern
- Grafikanalysen und GNN mit DGL und PyG
- Verteilte Trang/Inferenz auf Big-DL
- LLM-Vorab-Training und Feinabstimmung auf Ray

Tools

- Intel® Distribution for Python
- Intel® Optimized Modin
- Intel® Optimized XGBoost
- Intel® Extension for Scikit-Learn
- Intel® Optimized Tensorflow (ITEX)
- Intel® Optimized PyTorch (stock & IPEX)
- Intel® Neural Compressor
- SigOpt Python SDK und CLI
- CNVRG Python SDK und CLI
- Intel® Optimized Horovod
- DeepSpeed

Domänen-Kits

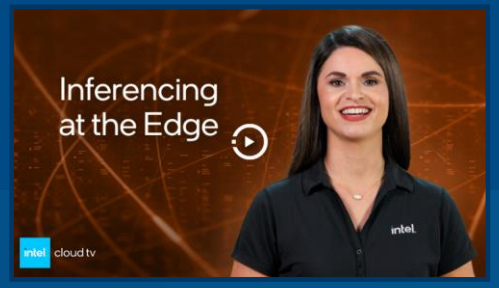
- Zeitreihen
- PPML
- Wissensübertragung
- Transformator/NLP

Die **Referenzkits** werden als **Container** geliefert und können in **gängigen Clouds sowie lokal verwendet werden.** Die **Referenzkits** sind auf **Workflows** und **Domänen-Toolkits** aufgeschichtet, die unabhängig voneinander genutzt werden können, um eine **breitere Auswahl von Anwendungsfällen in mehreren Branchen** zu unterstützen.

Cloud TV

Intel® Cloud TV berichtet über Cloud-Computing Nachrichten, Trends und Strategien, um Ihren Erfolg zu fördern

Ihre GenAI-Möglichkeit mit Intel® Gaudi® KI-Beschleunigern



Gewinnen Sie Erkenntnisse mit Dateninferenzierung am Edge

Mit KI in der Cloud Wettbewerbsvorteile schaffen

KI-Inferenzierung mit Cloud-Technologien

KI in der Cloud

Beschleunigen Sie die Skalierung von KI überall

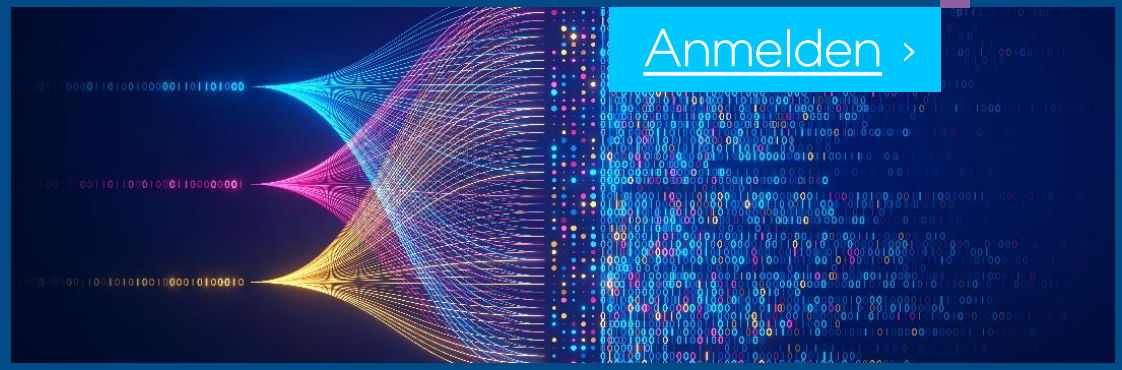
Schulung

KI überall verfügbar machen – Generative KI für Unternehmen (Anwendungsfälle)

Generative KI ist nicht nur für Internet-Chatbots. Unzählige Unternehmen überlegen, wie sie die Leistung von generativer KI und Large Language Models nutzen können, um den täglichen Betrieb zu unterstützen. In dieser Sitzung werden die Anwendungsfälle für generative KI in Unternehmen untersucht und Überlegungen dazu angestellt, wie Ihr Unternehmen sie im täglichen Betrieb einsetzen könnte.

Anmelden ›

Optimieren Sie KI für die Datengenerierung und Large Language Models



Anmelden ›

Die Integration von KI in die Workloads eines Unternehmens und die Skalierung einer bereits bestehenden Infrastruktur ist kompliziert und rechenintensiv. Sie erfordert die Entwicklung robuster Modelle, die auf massiven Datenmengen trainiert werden, und leistungsstarke GPUs, auf denen sie ausgeführt werden können. Nicht jedes Unternehmen hat die erforderlichen Ressourcen für so ein Unterfangen.

Diese Sitzung befasst sich mit einer Lösung: Eine Sammlung von Open-Source-KI-Referenzkits von Accenture* und Intel, die Unternehmen den Zugang zu KI erleichtern und für verbesserte Trainings- und Inferenzzeiten optimiert sind.

Zusätzliche Schulungen

Technisch

Ressourcentyp	Titel und Link
Kompetenz	KI in der Cloud – Kompetenz
Webinar	Optimierung von KI für Intel® Hardware mit Hugging Face
Webinar	Einrichtung von Cloud-basiertem verteiltem Training zur Feinabstimmung eines LLM
Schulungskurs	Verbesserung von LLMs mit Prompt Economization und In-Context-Learning
Schulungskurs	Optimieren Sie KI für die Datengenerierung und Large Language Models
Schulungskurs	Natürliche Sprachverarbeitung
Schulungskurs	Angewandtes Deep Learning mit TensorFlow*
Schulungskurs	Klein und flexibel – der schnelle Weg zu Enterprise GenAI
Schulungskurs	Die nächste Welle von GenAI – domänenspezifische LLMs
Leitfaden	Ein Entwickler-Leitfaden für die ersten Schritte mit generativer KI: Ein anwendungsspezifischer Ansatz
Schulungskurs	KI auf Intel® Xeon® Prozessoren in den Lösungsbereich bringen

Zusätzliche Schulungen

Nicht technisch

Ressourcentyp	Titel und Link
Video-Serie	Nutzung generativer KI
Schulungskurs	Klein und flexibel – der schnelle Weg zu Enterprise GenAI
Schulungskurs	Die nächste Welle von GenAI – domänenspezifische LLMs
Schulungskurs	Kompetenz „Prinzipien von KI überall“
Schulungskurs	Kompetenz „Prinzipien von KI-Software und -Technologieumfeld“
Schulungskurs	Nutzung des KI-Ökosystems: Mit Software gewinnen, mit SIs skalieren und die Lösung verkaufen
Schulungskurs	Generative KI und Large Language Models für die reale Welt

Weitere Ressourcen

Ressourcentyp	Titel und Link
Webinar	Generative KI-Webinar-Reihe
Webinar	GenAI überall verfügbar machen
Podcast	Wie Copilot, ChatGPT, Stable Diffusion und generative KI die Art und Weise verändern werden, wie wir entwickeln, arbeiten und leben
Beschreibung für Unternehmen	KI überall bereitstellen
Blog-Reihe	Tuning und Inferenz für generative KI mit Intel Xeon Prozessoren der 4. Generation
Lösungsbeschreibung	Bereitstellung und Skalierung generativer KI-Inferenz mit Lenovo ThinkSystem SR650 V3/Intel Xeon Prozessoren der 4. Generation Neue Intel und VMware Technik befeuern Lenovo ThinkAgile VX V3 Systeme
Tech-Artikel	Beschleunigung von Llama 2 mit Intel® KI-Hardware- und Software-Optimierungen
Forschungs-PR	10 % der befragten Unternehmen haben 2023 GenAI-Lösungen für die Produktion eingeführt
Video – Gesprächsrunde	Annahme der Rechen- und Nachhaltigkeitsherausforderungen der generativen KI
Podcast	Hugging Face und Intel – der Weg zu praktischen, schnelleren, demokratisierten und ethischen KI-Lösungen
Twitter/X-Gespräch	Wie demokratisierte große Sprachmodelle die KI-Entwicklung fördern
Supermicro-Benchmarks	Habana Validierung von Ansprüchen
Hugging-Face-Benchmarks	Benchmarks
Schulung/Webinar	Cloud Solution Architect (CSA) Tech-Talk: KI mit Habana
Whitepaper	Bei KI in Unternehmen dreht sich alles um das Entwickler-Whitepaper
Infografik	CPUs sind der Schlüssel für KI-Lösungen

Weitere Ressourcen

Ressourcentyp	Titel und Link
Lösungsbeschreibung	Optimieren Sie die KI-Einführung und -Bereitstellung mit Intel Enterprise AI mit Red Hat® OpenShift® AI
Leitfaden	Der KI-Leitfaden
Referenzkit	KI-unstrukturierte Textdatengenerierung
Whitepaper	Zoho optimiert und beschleunigt Video-KI-Workloads
Whitepaper	Seekr entwickelt ein vertrauenswürdiges KI-Screening-System
Lösungsbeschreibung	Sicherheit im Bildungsbereich: KI und Confidential Computing helfen dabei, sichere Remote-Prüfungen zu ermöglichen
Fallstudie und Video	Nature Fresh Farms nutzt KI vom Seed bis zum Store
Fallstudie	QMed Asia steigert die Krebserkennungsrate im Frühstadium
Fallstudie und Video	MetaApp erneuert KI-basiertes Empfehlungssystem
Lösungsbeschreibung	Optimierung der KI-Modell-Schulung und -Verfeinerung für die automatisierte optische Inspektion (AOI)
Blog	Prompt-gesteuerte Effizienzen für LLMs

Rechtshinweise und Disclaimer

Hinweise und Disclaimer.


intel