Gemeinsam die Zukunft der KI gestalten

Modernste Suchtechnologien für Spitzenleistungen in der KI-Entwicklung

Durchbruch in der KI

Jina AI und die Revolution der Informationsbeschaffung

Das Problem

In einer Ära des technologischen Fortschritts und der ständig steigenden Nachfrage nach Innovationen stehen KI-Entwickler vor anspruchsvollen Herausforderungen. Maximilian Werk, Head of Engineering bei Jina AI, einem Unternehmen, das sich darauf spezialisiert hat, die Leistung neuronaler Suchen zu verbessern, kennt diese Herausforderungen aus erster Hand. Mit einem Hintergrund in Mathematik und jahrelanger Erfahrung im Bereich des maschinellen Lernens, bringt er ein tiefgreifendes Verständnis für die Bedürfnisse seiner Branche mit.

Jina AI strebt danach, die Entwicklung von KI-Modellen zu erleichtern, die in verschiedenen Anwendungsbereichen eingesetzt werden können. Doch wie viele Unternehmen stand auch Jina AI vor der Herausforderung des Zugangs zu leistungsfähiger Hardware, insbesondere GPUs. Die Entscheidung, GPUs zu kaufen statt zu mieten, war ein bedeutender Schritt, der jedoch spezifische Hindernisse mit sich brachte.

Forschungsschwerpunkte

Jina AI, ein Berliner Start-up, gegründet 2020, ist ein führender Akteur in der künstlichen Intelligenz.
Der Schwerpunkt des Unternehmens liegt auf der Verbesserung der Internetsuche, insbesondere durch das Retrieval Augmented Generation (RAG)-System, das die Such- und Nutzungsmethoden von Informationen grundlegend verändern soll.

Seit seiner Gründung strebt Jina AI danach, die Leistung neuronaler Suchen zu verbessern, indem es sowohl Embedding- als auch Reranking-Modelle entwickelt. Diese sollen die Suche nach relevanten Informa-tionen erleichtern und die Reihenfolge der Suchergebnisse optimieren.

Anwendungsgebiete erklärt

Neural Search

Eine intelligente Suchfunktion, die relevante Informationen aus großen Datensätzen extrahiert, ähnlich einem persönlichen Assistenten, der genau das findet, wonach du suchst.

Deduplication

Eine Duplikaterkennungstechnologie, die doppelte Dateien oder Inhalte aufspürt und entfernt, um Ordnung zu halten und Speicherplatz zu sparen, vergleichbar mit einem digitalen Detektiv.

Classification

Eine automatische Klassifizierungsfunktion, die Inhalte in verschiedene Kategorien sortiert, ähnlich dem automatischen Sortieren von E-Mails in verschiedene Ordner basierend auf ihrem Inhalt.

Mit diesen Schwerpunkten steht Jina AI an der Spitze der KI-Entwicklung und treibt kontinuierlich Innovationen voran.

Die Hauptkunden von Jina AI sind Anwendungsentwickler, insbesondere für RAG-Programme, die moderne Chatbots ermöglichen.

Was bedeutet RAG?

Was ist Retrieval Augmented Generation (RAG)?

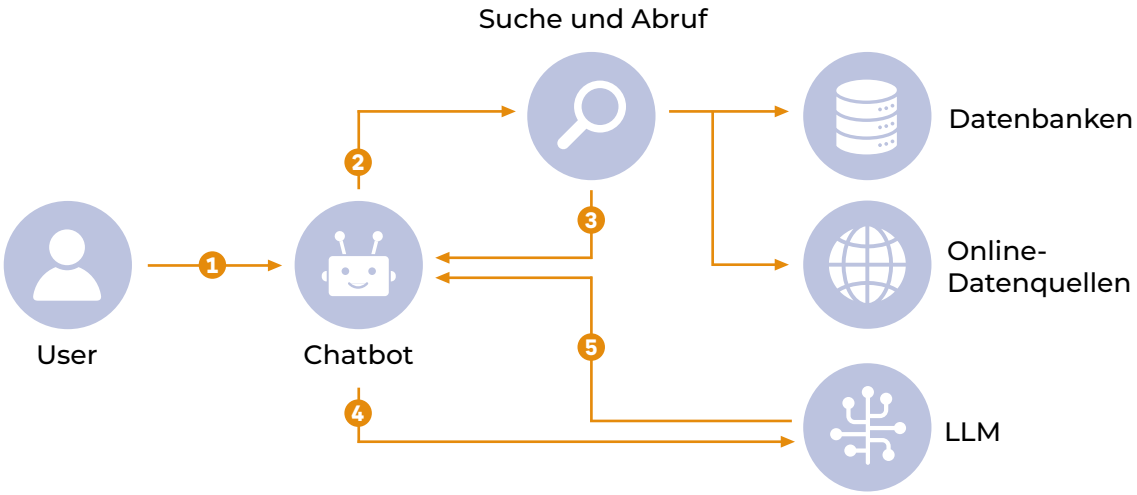
RAG kombiniert die Informationssuche mit künstlicher Intelligenz, um präzise Antworten zu liefern. Diese Technologie wird oft genutzt, um große Datenmengen effizient zu durchsuchen und spezifische, fundierte Antworten zu generieren.

Ist RAG dasselbe wie generative KI?

Nein, RAG ist eine Technik, die präzisere Ergebnisse für Abfragen liefern kann als ein generatives Sprachmodell allein. Während generative KI auf den in ihr trainierten Daten basiert, nutzt RAG zusätzlich externes Wissen, um genauere Antworten zu geben.

Wie wird RAG von generativer KI genutzt?

Unternehmensdaten werden in ein Wissensrepository eingebettet und in Vektoren umgewandelt, die in einer Vektordatenbank gespeichert werden. Bei einer Anfrage ruft die Vektordatenbank relevante Kontextinformationen ab. Diese Informationen werden zusammen mit der Anfrage an das große Sprachmodell gesendet, das den Kontext nutzt, um eine zeitnahe, präzisere und kontextbezogenere Antwort zu erstellen.



Vorteile von RAG? ↓

Was sind die Vorteile von RAG?

Die Retrieval-Augmented Generation (RAG) bietet zahlreiche Vorteile, um die Qualität der Reaktionen eines generativen KI-Systems auf Eingabeaufforderungen zu verbessern. Diese Vorteile gehen über die Möglichkeiten eines großen Sprachmodells (LLM) hinaus und umfassen:

Aktualität der Informationen

RAG hat Zugriff auf potenziell aktuellere Daten als die, die zur Schulung eines LLM verwendet wurden. Dies reduziert das Risiko von „Halluzinationen“ – falschen oder veralteten Antworten aufgrund unzureichender Informationen.

Kontinuierliche Aktualisierung

Das Wissensrepository von RAG kann kontinuierlich aktualisiert werden, ohne dass erhebliche Kosten anfallen. Dies gewährleistet, dass die Informationen stets auf dem neuesten Stand sind.

Kontextualisierung

Die Daten im Wissensspeicher von RAG können spezifischer und kontextbezogener sein als die in einem generalisierten LLM, was zu präziseren Antworten führt.

Fehlerkorrektur

Die Quellen der Informationen in der Vektordatenbank von RAG sind identifizierbar, was es ermöglicht, fehlerhafte Angaben zu korrigieren oder zu löschen. Dadurch wird die Zuverlässigkeit der generierten Antworten erhöht.

Leitplanken

Es können Leitplanken implementiert werden, um sicherzustellen, dass die generierten Antworten den gewünschten sprachlichen Stil beibehalten und unangemessene oder fehlerhafte Inhalte vermieden werden.

Die Herausforderungen

Von Miete zu Eigentum

Strategischer Wechsel zu eigenen GPUs

Die Anforderungen an KI-Entwickler wie Jina AI sind somit anspruchsvoll und vielfältig, besonders in einer Zeit des rasanten technologischen Fortschritts und steigender Nachfrage nach Innovationen und immer höheren Standards. Der Zugang zu leistungsstarker Hardware wie den GPUs ist dabei von entscheidender Bedeutung.

Die bisherige Methode, GPUs zu mieten, war für Jina AI nicht ausreichend und langfristig nicht tragbar. Die Mietkosten waren besonders bei voller Auslastung über mindestens 12 Monate viel höher als die Kosten für den Kauf. Zudem gab es oft Probleme, bestimmte GPU-Modelle kurzfristig zu mieten.

Deshalb entschied sich Jina AI, die GPUs zu kaufen. Dies war nicht nur wirtschaftlicher, sondern auch strategisch besser für die langfristige Planung und Kosteneffizienz.

Beim Kauf der GPUs stand Jina AI vor einer großen Herausforderung: Die gewünschten NVIDIA H100 GPUs waren schwer zu bekommen. Außerdem fehlte es an Erfahrung in der Einrichtung und Wartung eigener Hardware. Um diese Hürden zu überwinden, holte sich Jina AI die sysGen an Bord. sysGen ist ein Experte im Hosting und in der Konfiguration von High-Performance-Hardware.

Innovative IT-Partnerschaft

Jina AI und sysGen

sysGen ist eine Bremer Firma – ansässig im schönen Norddeutschland. Ausgestattet mit der neusten Technologie und angetrieben von einem hochmotivierten und internationalen Team vieler Nationalitäten, was immer auf dem neusten Stand ist. Nach der Prämisse – Arbeit muss Spaß machen – arbeitet sysGen wirkungsvoll im Team mit viel Motivation und einem respektvollen Umgang miteinander.

Als lösungsorientierter und herstellerunabhängiger IT-Ausrüster für Industrie, Handel, Forschung und Lehre ist sysGen darauf spezialisiert, hochwertige Systemlösungen, Server, Workstation und Komponenten inklusive System- und Applikationssoftware und Dienstleistungen für den professionellen Einsatz zu entwickeln, zu produzieren und zu vertreiben.

Das umfangreiche IT-Angebot umfasst revolutionäre sowie zukunftweisende Lösungen für die Bereiche Artificial Intelligence (AI) / Deep Learning), High Performance Computing, High Availability Computing (HA), Software Defined Storage, -Network, Virtualisierung und Cloud Computing. Für alle Lösungen und Systemlieferungen werden zusätzliche Dienstleistungen wie Installation vor Ort mit Integration in vorhandene IT Infrastrukturen, Mitarbeiterschulungen für Hard- und Software angeboten. Den Angeboten wird auch eine qualifizierte Beratung vorgeschaltet, ein kostenpflichtiger, europaweiter IT Service rundet die Angebotspalette ab.

Als Partner renommierter Unternehmen wie NVIDIA, Supermicro, Gigabyte, Intel, DDN, GRAIDtech bietet sysGen innovative IT-Lösungen.

Die Partner auf einen Blick:

- Unternehmen: sysGen
  Hauptsitz: Bremen, Deutschland
  Partner: NVIDIA und Supermicro Elitepartner
  Branche: Multi-industry
  Spezialisierung: Entwicklung, Produktion und Vertrieb von hochwertigen Systemlösungen, Servern, Workstations und Komponenten

- Unternehmen: Jina AI
  Gründung: 2020
  Hauptsitz: Berlin, Deutschland
  Branche: Softwareentwicklung, Künstliche Intelligenz
  Spezialisierung: Multimodale KI, neuronale Suche

Die Lösung

Verfügbarkeit als Rettungsanker

Die Verfügbarkeit der NVIDIA GPUs war für Jina AI entscheidend und zugleich die größte Herausforderung. sysGen, als NVIDIA Elite Partner, ermöglichte den Zugang zu den NVIDIA H100 GPUs. Dank effizienter Lagerhaltung konnte sysGen die Systeme schnell bereitstellen, was für den Projekterfolg essenziell war.

Obwohl Jina AI anfangs keine Erfahrung mit Hosting und Hardwarekonfiguration hatte, erwies sich dies nicht als Hindernis. Da Jina AI keine Erfahrung mit eigenen Bare-Metal-Systemen hatte, erfolgte seitens sysGen eine umfangreiche Unterstützungsleistung bei der Vorinstallation der Software-Anwendungen.

Jina AI testete die Systeme zunächst bei sysGen, was ihnen die nötige Sicherheit gab, dass die Server im Zielrechenzentrum fehlerfrei laufen würden. Die Remote-Konfiguration erleichterte die Inbetriebnahme erheblich.

„sysGen hat immer schnell und unkompliziert geholfen. Wir haben von Anfang an klargestellt, dass wir keine Erfahrungen mit Bare-Metal-Servern haben und wurden dementsprechend gut beraten.“

Maximilian Werkv | Jin AI

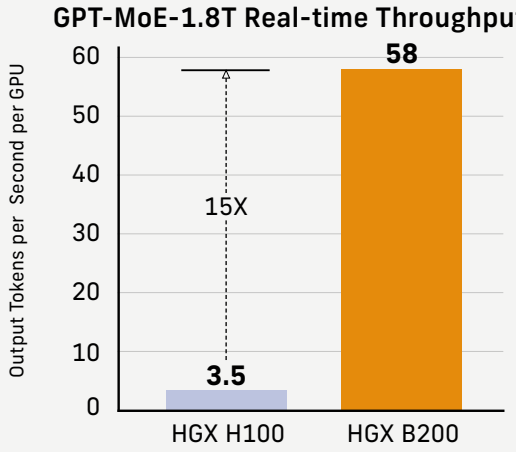
Das Ergebnis

Die weltweit führende KI-Computing-Plattform

Nach einer umfassenden Beratung entschied sich Jina AI für den Kauf von zwei High-Density-8U-Systemen von Supermicro (Modell HGX 8125GS-TNHR). Diese Systeme sind mit je acht NVIDIA® HGX™ H100 GPUs ausgestattet, die eine hohe GPU-Kommunikation durch NVIDIA® NVLINK™ und NVIDIA® NVSwitch™ ermöglichenund eignen sich perfekt eignen sich perfekt für High-Performance-Computing (HPC), Deep-Learning-Training, industrielle Automatisierung, Einzelhandel sowie Klima- und Wettermodellierung.

KI, komplexe Simulationen und massive Datensätze erfordern mehrere Grafikprozessoren mit extrem schnellen Verbindungen und einem vollständig beschleunigten Softwarestack. Die KI-Supercomputing-Plattform NVIDIA HGX™ vereint die volle Leistung von NVIDIA Grafikprozessoren, NVLink®, NVIDIA-Netzwerken und vollständig optimiertem KI- und High-Performance-Computing (HPC) Software-Stacks, um die höchste Anwendungsleistung bereitzustellen und die Zeit zum Erhalt von Einblicken so weit wie möglich zu verkürzen.

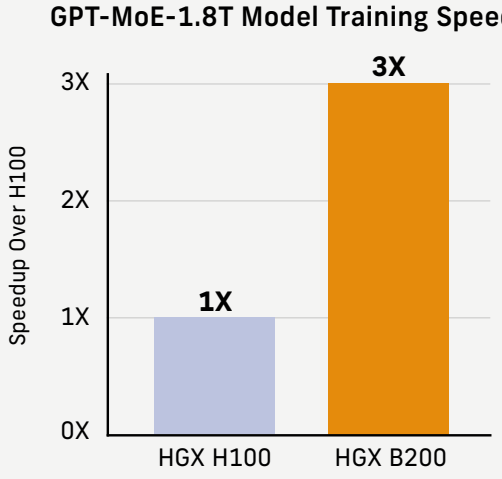
Deep-Learning-Inferenz: Leistung und Vielseitigkeit



Echtzeit-Inferenz für die nächste Generation großer Sprachmodelle

HGX B200 erzielt bei massiven Modellen wie dem GPT-MoE-1.8T eine bis zu 15-mal höhere Inferenzleistung als die vorherige Generation von NVIDIA Hopper™. Die Transformer-Engine der zweiten Generation verwendet individuelle Blackwell Tensor Core-Technologie in Kombination mit TensorRT™-LLM und Nemo™ Framework-Innovationen zur Beschleunigung der Inferenz für große Sprachmodelle (LLMs) und Mixture-of-Experts(MoE)-Modelle.

Deep-Learning-Training: Leistung und Skalierbarkeit



Trainingsleistung auf höchstem Niveau

Die Transformer Engine der zweiten Generation mit 8-Bit-Gleitkomma (FP8) und neuen Precisions ermöglicht bemerkenswerterweise ein 3-mal schnelleres Training für große Sprachmodelle wie GPT-MoE-1.8T. Dieser Durchbruch wird durch NVLink der fünften Generation mit 1,8 Terabyte pro Sekunde (TB/s) GPU-zu-GPU-Verbindung, InfiniBand-Netzwerke und NVIDIA Magnum IO™-Software unterstützt. Zusammen sorgen diese für effiziente Skalierbarkeit für Unternehmen und umfangreiche GPU-Computing-Cluster.

Quelle: https://www.nvidia.com/en-us/high-performance-computing/

Use Case

Das Herzstück im Einsatz

Durch den Einsatz der NVIDIA H100 GPUs von sysGen wurden hochmoderne Modelle trainiert, die die Relevanz der Suchergebnisse für einen Kunden, ein junges Startup-Unternehmen, erheblich verbesserten. Dieses Unternehmen unterstützt Vertriebsmitarbeiter dabei, während Telefonaten mit potenziellen Kunden genaue Antwortskripte zu erhalten. Mit den leistungsstarken GPUs konnten die Trainingszeiten drastisch verkürzt und die Genauigkeit der Modelle signifikant gesteigert werden. Dies ermöglichte es dem Startup-Unternehmen, schnellere und präzisere Antworten während der Kundeninteraktionen zu liefern, was zu einer verbesserten Kundenzufriedenheit und letztendlich zu einem erfolgreichen Geschäftsergebnis führte.

„Wir haben uns mit sysGen zusammengetan, um Embedding- und Reranking-Modelle für die RAG-Systeme unserer Kunden zu trainieren, und die Ergebnisse waren hervorragend.“

Maximilian Werk | Jina AI

RAG Anwendungsfall

Industrie-Analyse

Erstellung von Marktberichten mit RAG auf der Grundlage von Branchendaten.

Kundenbetreuung

Entwicklung von Chatbots für zuverlässige Hilfe, wie z. B. der Bot eines Einzelhändlers für Liefer- und Rückgaberichtlinien.

Inhaltserstellung

Einsatz von RAG für maßgeschneiderte Inhalte wie Artikel und Newsletter.

Assistenten für Dokumentenrecherche

Erstellung von Chatbots für HR-, Compliance- und Sicherheitsabfragen aus Unternehmensdokumenten.

Gesundheitsberatung

Bereitstellung medizinischer Informationen und Unterstützung durch RAG-gesteuerte Chatbots für 24/7- Patientenbetreuung.

Erfolgsbilanz

Jina AI und sysGen: Eine Erfolgsgeschichte

Dank der von sysGen bereitgestellten GPU-Systeme konnte Jina AI die Effizienz und Genauigkeit ihrer neuronalen Suchsysteme deutlich steigern. Der Erwerb eigener Hardware ermöglichte eine bessere Kosten-Nutzen-Kalkulation, da keine GPU-Leistung mehr gemietet werden muss. Diese Hardware zu besitzen, ermöglicht sowohl Kosten als auch Arbeitsprozesse zu optimieren. Zudem können die Trainingsdaten dauerhaft lokal gespeichert werden, was weniger organisatorische Arbeit und Zeitverschwendung beim Datentransfer bedeutet. Updates, Wartungen und Fehleranalysen erfolgen remote durch sysGen, was den organisatorischen Aufwand weiter reduziert.

Die Anschaffung von Bare-Metal-Systemen erwies sich nicht zuletzt dank der persönlichen Beratung im Vorfeld seitens sysGen als weniger kompliziert als gedacht.

„Für Modelltraining auf Single-Node GPU Instanzen ist die Lösung perfekt für uns. Ich vermute, dass wir im zweiten Halbjahr 2024 weitere GPU-Maschinen kaufen werden.“

Maximilian Werk | Jina AI

Möchten auch Sie Ihre IT-Infrastruktur optimieren und von maßgeschneiderten Lösungen profitieren? Kontaktieren Sie Ihre Ansprechpartner bei sysGen und erfahren Sie mehr über die individuellen Beratungs- und Serviceangebote der sysGen.

Ihre Ansprechpartner:

Gabriele Nikisch

Geschäftsführung / Vertriebsleitung

Tel: 0421 409 66 21

gnikisch@sysgen.de

Sergius Siczek

Geschäftsführer / Technikleitung

0421 409 66 32

ssiczek@sysgen.de

sysGen GmbH, Am Hallacker 48, 28327 Bremen

Besuchen Sie unserer Website:

COMING SOON!

Neu!

NVIDIA DGX B200

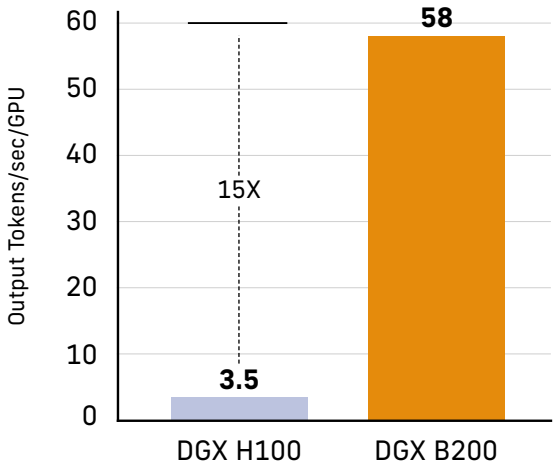
Die Grundlage für Ihr KI-Kompetenzzentrum.

NVIDIA DGX™ B200 ist ausgestattet mit acht NVIDIA Blackwell-GPUs, die über NVIDIA® NVLink® der fünften Generation miteinander verbunden sind, bietet DGX B200 eine bahnbrechende Leistung mit der 3-fachen Trainingsleistung und der 15-fachen Inferenzleistung der Vorgängergenerationen. Durch die Nutzung der NVIDIA Blackwell-GPU-Architektur kann DGX B200 diverse Workloads bewältigen, einschließlich großer Sprachmodelle, Empfehlungssysteme und Chatbots, und ist damit ideal für Unternehmen geeignet, die ihre KI-Transformation beschleunigen möchten.

NVIDIA DGX™ B200: Die Einheitliche KI-Plattform für Unternehmens-Pipelines von Entwicklung bis Bereitstellung

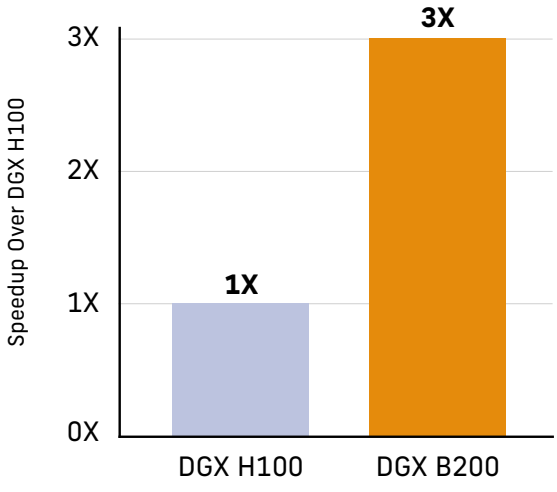
NVIDIA DGX-Systeme setzen den Standard für KI-Leistung und Effizienz. Mehr Information.

Echtzeit-Inferenz für Large Language Models



GPT-MoE-1.8T Real-time Throughput

Neue Maßstäbe bei der KI-Trainingsleistung



GPT-MoE-1.8T Model Training Speed-Up

Quelle: https://www.nvidia.com/en-us/data-center/dgx-b200/#referrer=vanity