

DIZ KI Strategie und ihre Umsetzung in der Universitätsmedizin

Dr. Tim Herrmann

21.10.2025

Kolloquium

DIZ und KI ?

- Was hat ein DIZ mit KI zu tun?
- Ein DIZ verfügt über Daten die nicht Teil der LLM sind
 - **Dies bietet Chancen !**
- Dafür muss ein DIZ selbständig mit KI Systemen arbeiten können und diese vollständig kontrollieren

DIZ
mit
KI

KI für Forschung und Lehre

- Exponentiell wachsende Bedeutung
- Hat Auswirkung auch auf die Struktur einer Universität
 - z.B. Bedarf für Übersetzungspersonal wird verschwinden
- Herausforderung KI nachhaltig und wettbewerbsfähig in Forschung und Lehre zu integrieren

nature medicine

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature medicine](#) > [brief communications](#) > [article](#)

Brief Communication | [Open access](#) | Published: 23 April 2025

Benchmark evaluation of DeepSeek large language models in clinical decision-making

[Sarah Sandmann](#), [Stefan Hegselmann](#), [Michael Fujarski](#), [Lucas Bickmann](#), [Benjamin Wild](#), [Roland Eils](#)✉ & [Julian Varghese](#)

[Nature Medicine](#) (2025) | [Cite this article](#)

6919 Accesses | 46 Altmetric | [Metrics](#)

Abstract

Large language models (LLMs) are increasingly transforming medical applications. However, proprietary models such as GPT-4o face significant barriers to clinical adoption because they cannot be deployed on site within healthcare institutions, making them noncompliant with stringent privacy regulations. Recent advancements in open-source LLMs such as DeepSeek models offer a promising alternative because they allow efficient fine-tuning on local data in hospitals with advanced information technology infrastructure. Here, to demonstrate the clinical utility of DeepSeek-V3 and DeepSeek-RL, we benchmarked their performance on clinical decision support tasks against proprietary LLMs, including GPT-4o and Gemini-2.0 Flash Thinking Experimental. Using 125 patient cases with sufficient statistical power, covering a broad range of frequent and rare diseases, we found that DeepSeek models perform equally well and in some cases better than proprietary LLMs. Our study demonstrates that open-source LLMs can provide a scalable pathway for secure model training enabling real-world medical applications in accordance with data privacy and healthcare regulations.

Fragestellungen:

- Anwendungsfälle
- Cloud vs. OnPrem KI
 - OpenSource vs. geschlossene Systeme
 - Vor- und Nachteile
- Bei OnPrem
 - welche Hardwarestrategie
- Wie umgehen mit KI Agents und Connectoren
 - Große Potenziale und möglicher Missbrauch
- Qualität und Standards von KI

KI Anwendungsfälle:

- Datenanalyse
 - Recherchen
- Datengenerierung
- Dokument arbeiten
 - Texte erstellen, übersetzen
umformulieren
- Gutachten Erstellung und
Auswertung
- Simulationen
- Systemsteuerung

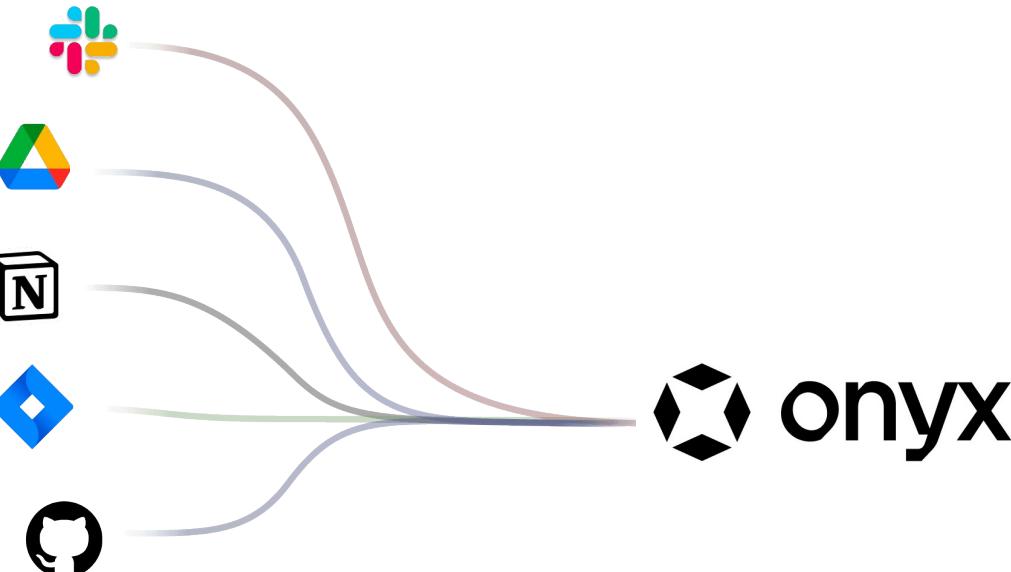
- Antragstexte in Projektskizzen
und AZA Bögen
 - kürzen oder erweitern

Cloud vs. OnPrem KI

- Cloud langfristig teurere Alternative
- Cloud erzeugt Abhängigkeit und führt häufig zu Vendor Lock
- Cloud sinnvoll bei Internet gebundenen Dienstleistungen, z.B. Zoom
- OnPrem personal intensiver
 - muss kontinuierlich weiterentwickelt und an die lokalen Bedürfnisse angepasst werden

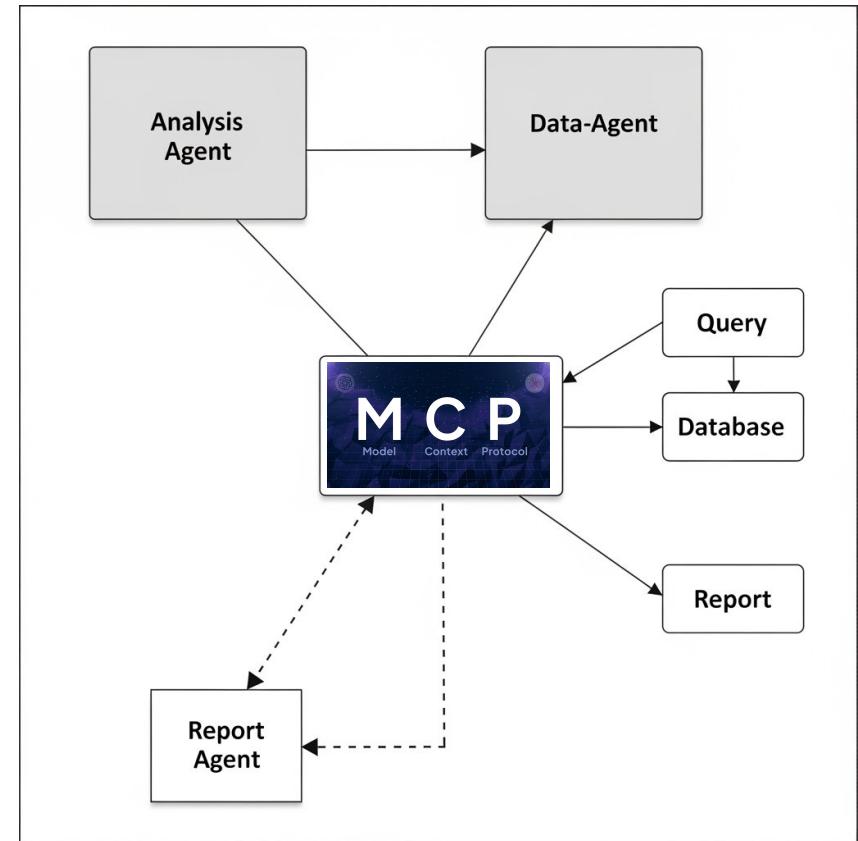
KI Connector und Assistenzsystems

- Onyx KI Connector:
 - So wird eine zentrale Anbindung und Analyse aller User Spaces ermöglicht: Nextcloud, Confluence, 3rd Party Systeme
 - <https://github.com/onyx-dot-app/onyx>



KI Connector und Assistenzsystems

- Model Context Protocol (MCP)
- offener Standard und Open-Source-Framework
- um die Integration und den Datenaustausch zwischen KI und externen Tools, Systemen sowie Datenquellen zu standardisieren
- Grundlage für KI Agenten bzw. Assistenzsysteme



Qualität von KI

- Es ist schwer, in den vielen Bereichen die Qualität von KI zu beurteilen,
- Es gibt Bereiche die aber sehr Allgemein für viele erahnen lassen, wie gut die Qualität sich innerhalb von kürzester Zeit verbessert

Qualität von KI

■ Benchmarks

Category Benchmark	Llama 4 Scout	Llama 3.3 70B	Llama 3.1 405B	Gemma 3 (27B)	Mistral 3.1 (24B)	Gemini 2.0 Flash-Lite
Image Reasoning MMMU	69.4	No multimodal support	No multimodal support	64.9	62.8	68.0
	70.7			67.6	68.9	57.6
Image Understanding ChartQA	88.8	No multimodal support	No multimodal support	76.3	86.2	73.0
	94.4			90.4	94.1	91.2
Coding LiveCodeBench (10/01/2024-02/01/2025)	32.8	33.3	27.7	29.7	-	28.9
Reasoning & Knowledge MMLU Pro	74.3	68.9	73.4	67.5	66.8	71.6
	57.2	50.5	49.0	42.4	46	51.5
Long context MTOB (half book) eng->kgv/kgv->eng	42.2/36.6	Context window is 128K	42.3/35.1 ³			
	39.7/36.3					35.1/30 ³

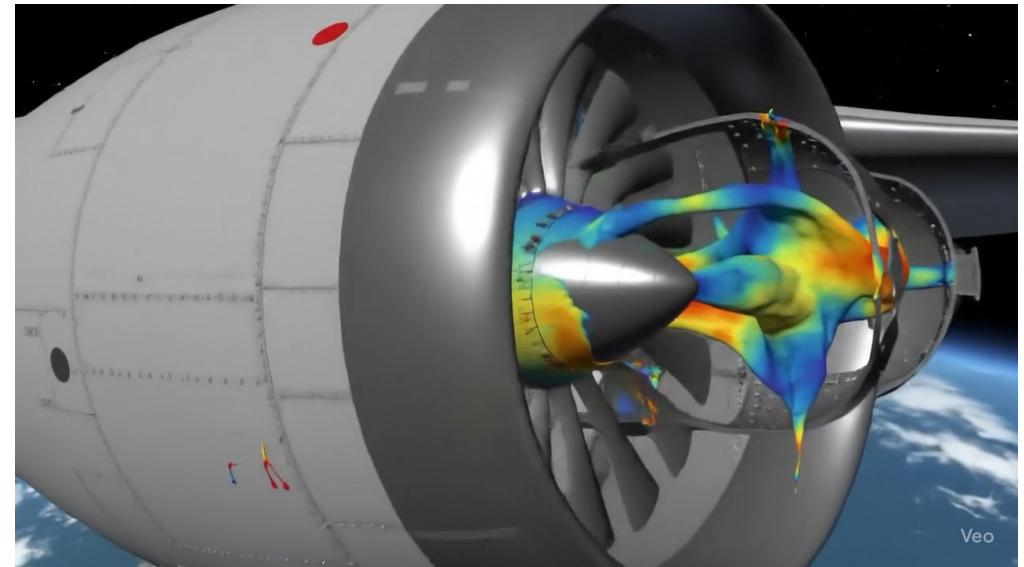
1. For Llama model results, we report 0 shot evaluation with temperature = 0 and no majority voting or parallel test time compute. For high-variance benchmarks (GPQA Diamond, LiveCodeBench), we average over multiple generations to reduce uncertainty.
2. For non-Llama models, we source the highest available self-reported eval results unless otherwise specified. We only include evals from models that have reproducible evals (via API or open weights), and we only include non-thinking models
3. Specialized long context evals are not traditionally reported for generalist models, so we share internal runs to showcase llama's frontier performance.

Qualität von KI

- Video und Bildergenerierung



 OpenAI
Sora



Triebwerk Strömungssimulation mit Veo 2 generiert

Softwareestrategie bei OnPrem:

- Ollama und OpenWebUI
- HAWKI



User Interface HAWKI vs. OpenWebUI bei OnPrem:

HAWKI

- Seit 06/2023
- 16 Entwickler
- Anbindung:
 - Cloud: OpenAI
 - OnPrem: LLM
- Weitere Features:
 - User Anonymisierung zu ChatGPT

OpenWebUI

- Seit 02/2024
- 564 Entwickler
- Anbindung:
 - Cloud: OpenAI
 - OnPrem: LLM
- Weitere Features:
 - Audio Transkription
 - Model Builder
 - MCP
 - Image Generation Integration

Mögliche Hardwarestrategie bei OnPrem:

Ebenen

- Universitätsmedizin
 - GPU Server 4xH200 0,2 Mio € **seit 10/2025**
- Universität
 - GPU Server 8x Nvidia B300 0,5 Mio €
- Bundesland
 - GPU Cluster für Scale-up (Erweiterung im Rack) oder Scale-out

Beispiel Sachsen-Anhalt - Rohbau des IT-Knotenpunkts für die Hochschulen 10/2025



Option AMD Helios - AMD Advancing AI Keynote 12.06.2025

Previewing Today at **Advancing AI 2025**

AMD “Helios” Optimized AI Rack Solution

AMD EPYC | AMD INSTINCT | AMD PENSANDO | AMD ROCm

Available in 2026



OPEN Compute Project | ULTRA ACCELERATOR LINK™ | Ultra Ethernet

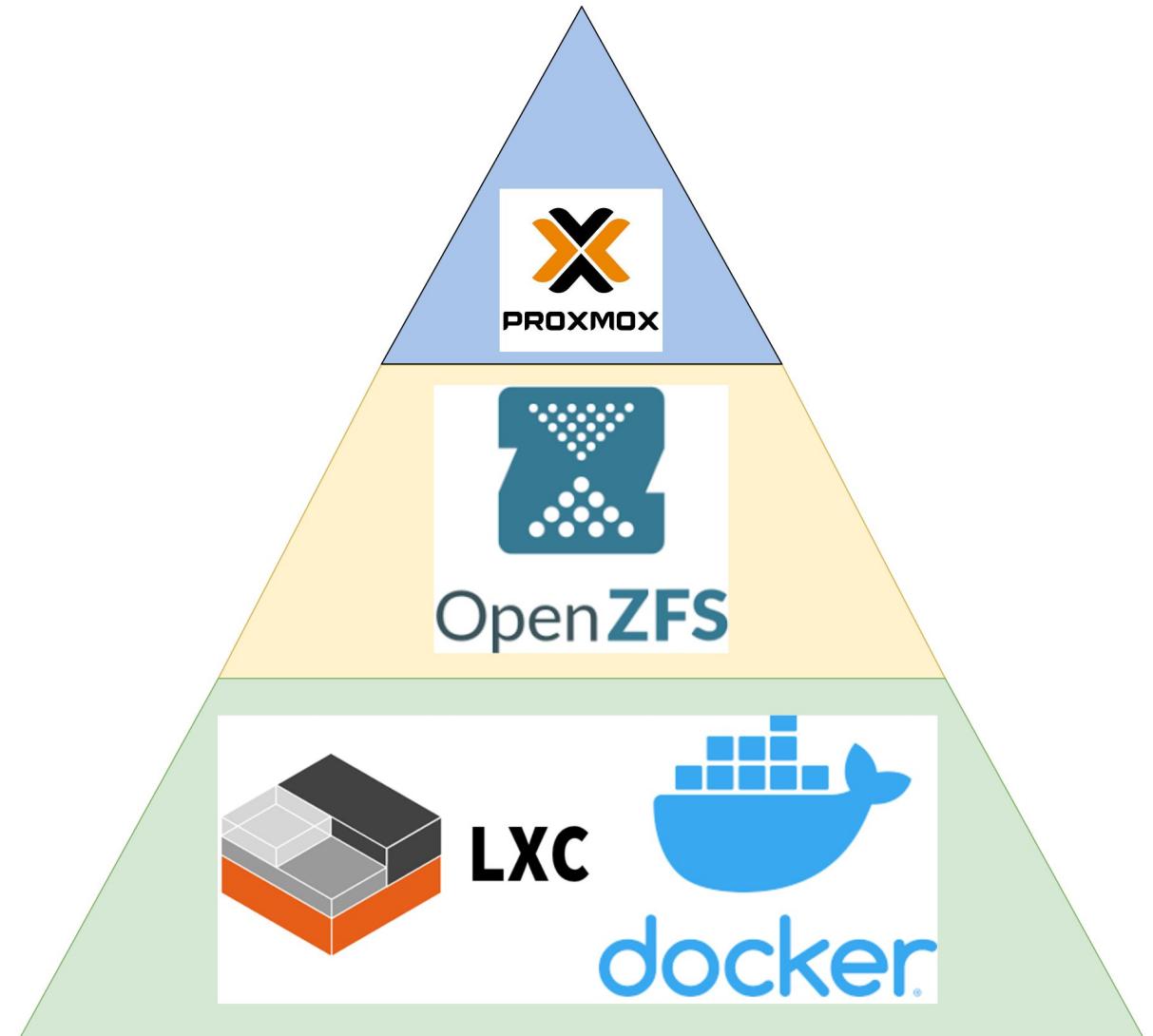
UMMD DIZ

KI Strategie

Virtualisierungskonzept - DIZ Magdeburg

Proxmox VE

- OpenZFS
- VM und LXC
- vGPU Konzept mit Proxmox VE
- **Vorteile:**
 - Geringe Kosten
 - Hoher Investitionsschutz



DIZ KI Plattform

- DIZ KI Plattform: <https://ki-plattform.diz-ag.med.ovgu.de>
- Angebunden an Keycloak Nutzerverzeichnisdienst
- Juli 2024 Testphase und seit **Jan. 2025** Produktivphase gestartet
- Über 400 User bis dato
- **Basis:** Open Source Software Komponenten:
 - Ollama Framework: <https://github.com/ollama/ollama>
 - Open WebUI: <https://github.com/open-webui/open-webui>
 - Stable Diffusion: <https://github.com/AUTOMATIC1111/stable-diffusion-webui>

DIZ KI Plattform

- **Hardware:**
 - Dual Socket Server
 - CPU: 2x AMD EPYC 9754 128-Core max. 3.10GHz
 - RAM: 2304GB (24x 96GB) DDR5-4800 DIMM, REG, ECC, 2R
 - Storage: 24x SSD 30.7TB U.3 Micron 9400 PRO
 - GPU: 1x NVIDIA H100, 80GB

>> GPU RAM limitiert den Einsatz für Größe der LLM <<

DIZ KI Plattform

The screenshot shows the DIZ KI Plattform interface. At the top, there is a header bar with the text "llama3.2-vision:90b" and a dropdown arrow, followed by a "Set as default" link and a user icon. Below the header is a navigation bar with "INFO KI Portal" and a close button "x". The main area features a dark background with a network graph pattern. In the center, there is a white rounded rectangle containing the text "llama3.2-vision:90b" next to a small "DIZ" logo. Below this, a search bar contains the placeholder "How can I help you today?". Underneath the search bar are three buttons: "+ Web Search", "Image", and "Code Interpreter". To the right of the search bar are two icons: a microphone and a speaker. Further down, there is a section titled "Suggested" with two items: "Erzähl mir einen Fun Fact über das römische Imperium" and "Hilfe bei Studie Größenvergleich".

DIZ KI Plattform inkl. Bildgenerierungsoption

Neue Unterhaltung = llama3.1:70b +

Arbeitsbereich

Suchen

Heute

Strand Sonnenaufgang

Vorherige 7 Tage

Data Donation for Research

Vorherige 30 Tage

Improving Medical Diagnosis

Gender-Sensitive Data Analyti

Beirat

Flow Chart Erstellung

Establishing Partner Collaborati

Docker Youtrack Setup

German Expression Gratitude

Sorted Schedules by Date

Keycloak & Apache Superset

liste mit komma getrennten Wörte

August

Keycloak Login Issue

Apache Superset

E=mc² Formel

Broad Consent Explained

Graph Database Optimizatio

MCAS Definition

Römisches Imperium Fun fact

Höchste Gebäude frühes 20. Jhd

Tim Herrmann

Generiere ein Bild von einem Sonnenaufgang am Strand.

llama3.1:70b 1:45 PM



Ich bin ein Textmodell, daher kann ich kein Bild generieren. Ich kann jedoch eine Beschreibung des Bildes erstellen:

Das Bild zeigt einen sonnigen Morgen am Strand. Die Sonne steigt langsam über dem Horizont auf und taucht den Himmel in ein warmes, goldenes Licht. Die Wellen des Meeres glänzen im Morgenlicht und reflektieren die Farben des Himmels. Der Sandstrand ist leer und still, mit nur wenigen Fußspuren, die sich bis zum Wasser hinziehen.

Im Vordergrund des Bildes liegt ein einsamer Strandkorb, der sanft im Wind schwankt. Die Stangen des Korbs sind mit Muscheln und anderen Meeressutensilien dekoriert. Hinter dem Korb erstreckt sich eine Düne, die von hohen Gräsern und wilden Blumen bedeckt ist.

Am Horizont kann man ein paar Segelboote erkennen, die langsam über das Wasser gleiten. Die Sonne steigt höher und höher, bis sie fast den gesamten Himmel ausfüllt. Das Bild strahlt eine friedliche und idyllische Atmosphäre aus und lädt den Betrachter ein, sich in die Stille des Morgens am Strand zu versetzen.

+ Eine Nachricht senden

Microphone icon

Headphones icon

DIZ KI Plattform

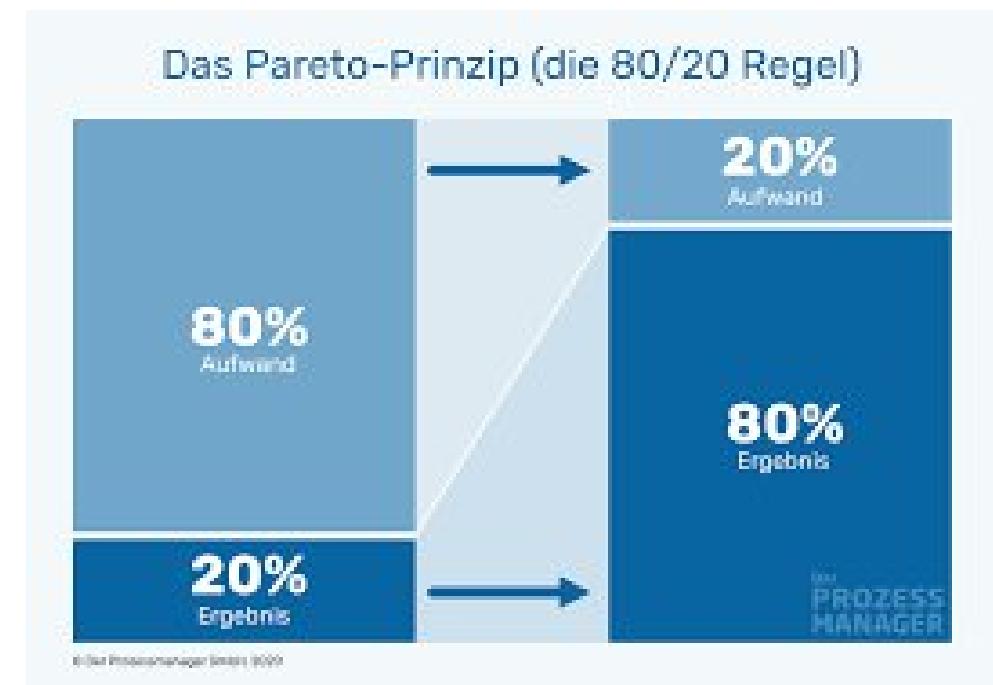
- Vergleich zu OpenAI ChatGPT

Pros:

- OnPrem alle Daten bleiben im eigenen Datennutzungsraum
- Keine Ausfälle durch Cloud Störfälle
(z.B. AWS Ausfall 20.10.2025)
- anwendungsbezogene Modelle

Cons:

- nur ca. 80% der Fähigkeiten und Datenqualität von den GPT 4
- Hardwareanforderungen werden steigen mit wachsenden Nutzerzahlen
- Pareto effizient ?



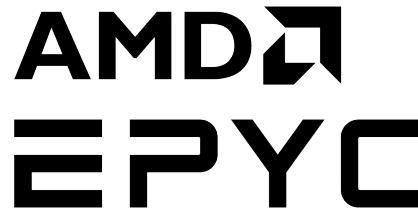
DIZ KI Plattform - Kostenkonzept

- Wie werden Kosten gemessen und verglichen?
 - gemessen in 1 Mio. **Token** Einheiten
 - Public Cloud Plattformen (OpenAI, Google Gemini) >> 10 bis 15 € pro 1 Mio. Output Token
 - Preise für die Input Token ist häufig 1/3 oder 1/5 davon
 - OnPrem **DIZ KI Plattform**:
 - 150 € für 1 Mio. für Output und Input Token
 - ca. Faktor 10x mehr Kosten
 - Kosten lassen sich senken
- Preis für Unabhängigkeit Gegenüber außer europäischen Anbietern
- Preis für Datensouveränität bei medizinischen Forschungsdaten und Patientendaten

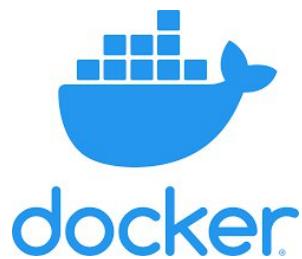
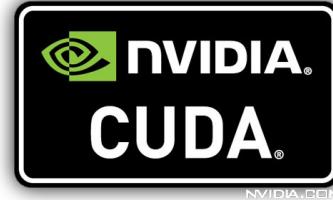
Umfrage zur Nutzung von DIZ KI-Plattform

- Umfrage an Nutzergruppe wurde durchgeführt
- Fragebogen umfasste:
 - Anwendungsfälle
 - Art der Daten die in die KI-Plattform übertragen werden
 - Zufriedenheit
 - Zuverlässigkeit

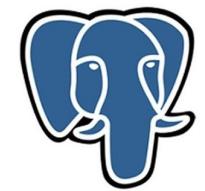
>> Veröffentlichung eingereicht



OpenZFS



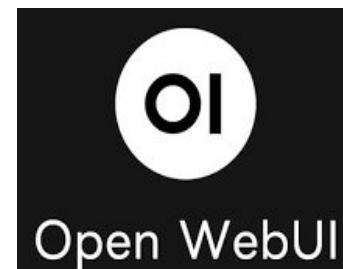
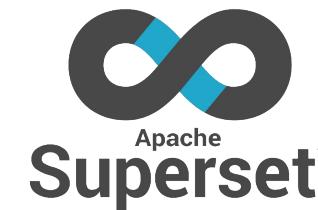
PROXMOX



PostgreSQL



Nextcloud



Diskussion

Kontakt

URL: <https://diz.med.ovgu.de>

E-mail: diz@med.ovgu.de

Folgen Sie uns auf LinkedIn



<https://linkedin.com/in/diz-datenintegrationszentrum-der-uni-medizin-magdeburg-96a355261>