

ChatAI - KI-Services in der Wissenschaft

Prof. Dr. Ing. Ramin Yahyapour
CIO Universität und
Universitätsmedizin Göttingen

Gesellschaft für wissenschaftliche
Datenverarbeitung

Stand 17.12.2024

Die GWWDG ist eine gemeinsame, gemeinnützige Einrichtung von Universität Göttingen und der Max-Planck-Gesellschaft

- Wissenschaftliches Rechenzentrum

Sie übernimmt auch diverse überregionale Aufgaben:

- Nationales Hochleistungsrechenzentrum
- Betreiber für die 7 norddt. Bundesländer
- Nationales HPC-Zentrum des DLR
- Nationales KI Servicezentrum
- Datenzentrum für verschiedene nationale Initiativen
- Cloud-Betreiber in der Wissenschaft, bspw. Academic Cloud für nds. Hochschulen

Zahlen und Fakten zu Services und Betrieb

- 150.000 Nutzende
- Gesamte Betriebskapazität (Strom) von > 5 MW
 - Verteilt auf 3 Rechenzentren
- Rechenzentrumsneubau aus 2022 bezogen
- Speichersysteme (Filesysteme, HSM, Backup, Archiv, ...) in der Größenordnung von 140 Petabyte
- 12.000 Server
- 30,5 Petaflops High Performance Computing
- Zertifiziert gemäß ISO 27001 (Informationssicherheit) und ISO 9001 (Qualitätsmanagement)
- Betreiber kritischer Infrastrukturen für sensible Daten³

KI-Servicezentrum für sensible und kritische Infrastrukturen



Eines von vier BMBF-geförderten nationalen KI Servicezentren

BMBF-gefördert: 20 Mio €

Ziel: Erforschung der Einrichtung eines KI-Servicezentrums

- Erfüllung der **Anforderungen** kritischer Infrastrukturen: Sicherheit, Datenschutz, Zuverlässigkeit
- Dienstleistungen für Pilotprojekte in ganz Deutschland
- Startups, KMUs können die Dienste **kostenlos** nutzen!

- **Forschung** zur weiteren Verbesserung der Dienstleistungen
Skalierbarkeit, Datenverwaltung, Portabilität
- Unentgeltliche Durchführung von Pilotprojekten
 - Beratungen, Rechenressourcen, Services, Kurse

4 Nationale KI Servicezentren

WestAI

Dortmund/Bonn/Jülich/Aachen/Paderborn

hessian AI Service Center
Darmstadt

KISSKI

Hannover/Göttingen/Kassel

KI-Servicezentrum Berlin Brandenburg

Hasso-Plattner-Institut

KISSKI Plattform

- Rechenressourcen für das Training großer, skalierbarer KI-Modelle
- Speicherkapazität für große Datenmengen
- Daten- und Modellkatalog
- Entwicklungsplattform

Schwerpunkt:
Lebenswissenschaften und Energie

- Zugang zu hochverfügbaren HPC-Ressourcen
- Beratung zur optimalen Nutzung von KI
- Entwicklungsleistungen für KI-Produkte
- Umfangreiches Schulungsangebot



ChatAI – Large Language Models

- Bedarf: Forschende & Studierende möchten LLMs nutzen
- 1) Angebot an kommerziellen externen Modelle (ChatGPT/OpenAI)
 - Einschränkungen von kommerziellen Anbietern
 - keine Datenschutzgarantie, Unkalkulierbare Kosten mit pay-as-you-go
- 2) Bereitstellung eines LLM-Service mit eigenen Modellen
 - Open Source Basis: 8 lokale Modelle
 - Leistungsstarke, skalierbarer Infrastruktur
 - Strenger Datenschutz, keine Speicherung von Prompts

Resultat: Chat AI

Chat The word "Chat" is in a large, light grey font. To its right is a stylized "AI" logo where the "A" is formed by a blue square and a purple triangle, and the "I" is a blue vertical bar.

<https://chat-ai.academiccloud.de/>

Academic Cloud Hub

Social network

The Academic Cloud Hub is a platform for communication and networking that brings together staff, lecturers and students from universities in Lower Saxony for project- or topic-related exchange.

[MORE INFO](#)

communication community teaching

Actionbound

Learning app

Actionbound promotes mobile learning through gamification. Interactive content such as quizzes, maps and tasks can be integrated into learning tours. Rankings and competitions increase motivation.

[MORE INFO](#)

teaching

BigBlueButton

Videoconference system

Digital tool for videoconferences with features such as breakout rooms, screen sharing, collaboration, chat, and whiteboard.

[MORE INFO](#)

videoconferencing

Chat AI

AI chatbot

Chat AI offers an easy and secure access to powerful generative AI. The intuitive interface allows users to chat directly with a selection of different AI models.

[MORE INFO](#)

AI teaching research

Chemotion

Electronic Lab Notebook

Chemotion ELN is a web-based application specifically designed for chemists. It provides a structured and efficient way to capture, organize, and analyze experimental data.

[MORE INFO](#)

research

Collaboard

Whiteboard

Collaboard is an interactive whiteboard for team collaboration. It enables users to collect ideas, plan complex projects visually and work together in real time.

[MORE INFO](#)

collaboration teaching whiteboard teams

ePIC

PID service

ePIC is a service for creating and editing unique persistent identifiers (PIDs) that can be attached to research data and documents, so that they can be

[MORE INFO](#)

GitLab

Source code management

GitLab is a web-based Git repository manager and facilitates software

[MORE INFO](#)

GRO.data

Data repository

Göttingen Research Online Data is an online repository for research data. It allows for data to be stored, edited and published along with metadata and identifiers (PIDs)

[MORE INFO](#)

Einbindung in
AcademicCloud

Für Niedersachsen, MPG und
nationale Initiativen

ACADEMIC TOOL BOX FOR YOUR RESEARCH, STUDY AND WORK

Seit 2015 eine Cloud für die Wissenschaft

Anwendungen

Sync and Share (ownCloud)

Gitlab

ShareLaTeX/Overleaf

DataVerse

BigBlueButton

PID-Service

Rocket Chat

OnlyOffice

JupyterNotebook (Im Testbetrieb)

Neu

Elektronisches Laborbuch

Knowledge-Graph

Collaboard

Hintergrunddienste

Portal

Dokumentationsserver

IDM + SSO

2FA (Im Testbetrieb)

ChatAI

Matrix

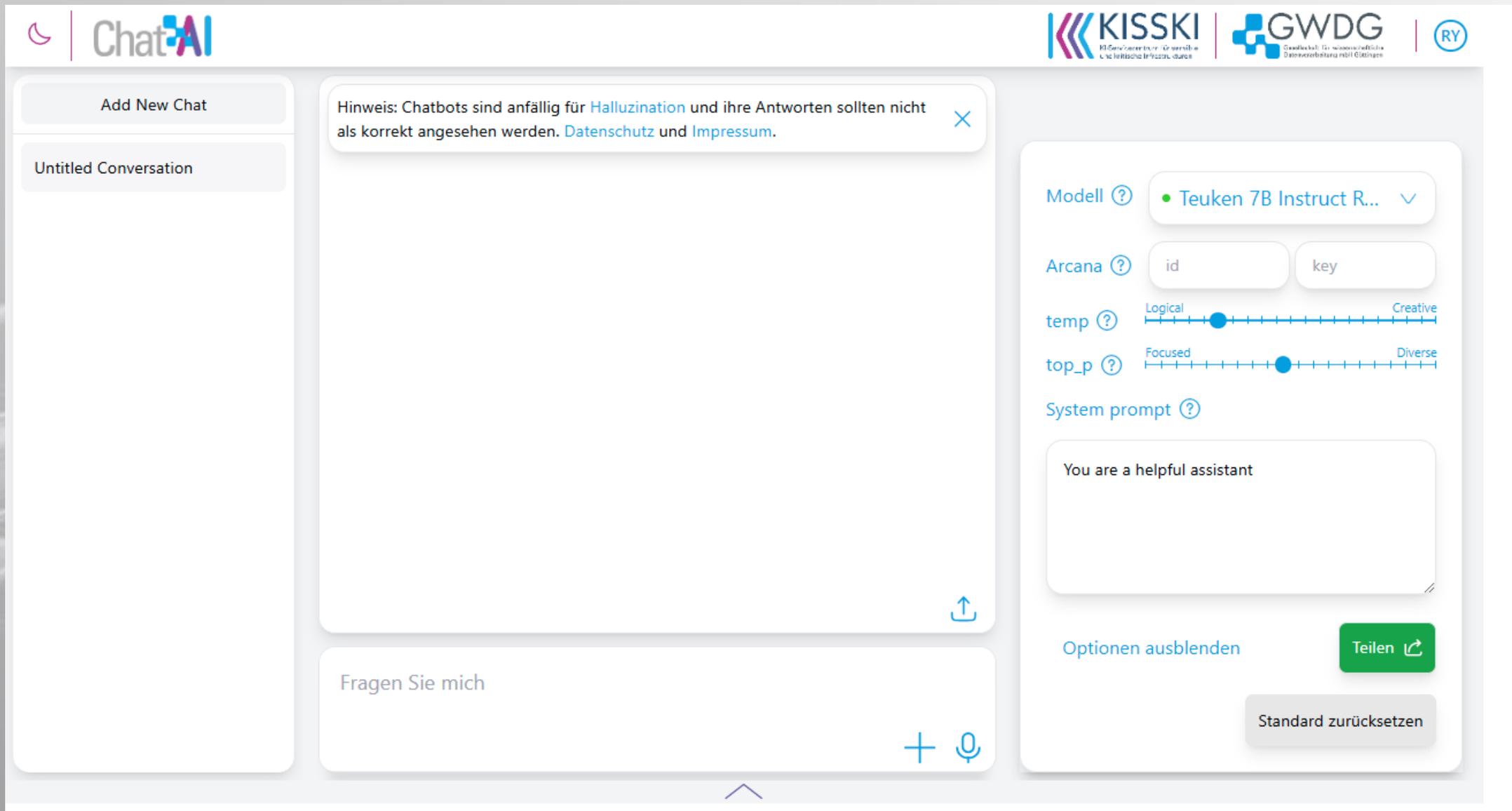
Academic Services for Niedersachsen

As a member of a participating university, research or higher education institute in Niedersachsen you can use the Academic Cloud Services. Log in and benefit from our expertise and service portfolio in the fields of storage, sharing, communication and useful working tools. For institutions outside of Niedersachsen, please

Just log in and get instant access to our tools:

- ✓ Share & store your work and LaTeX files
- ✓ Start and evaluate surveys

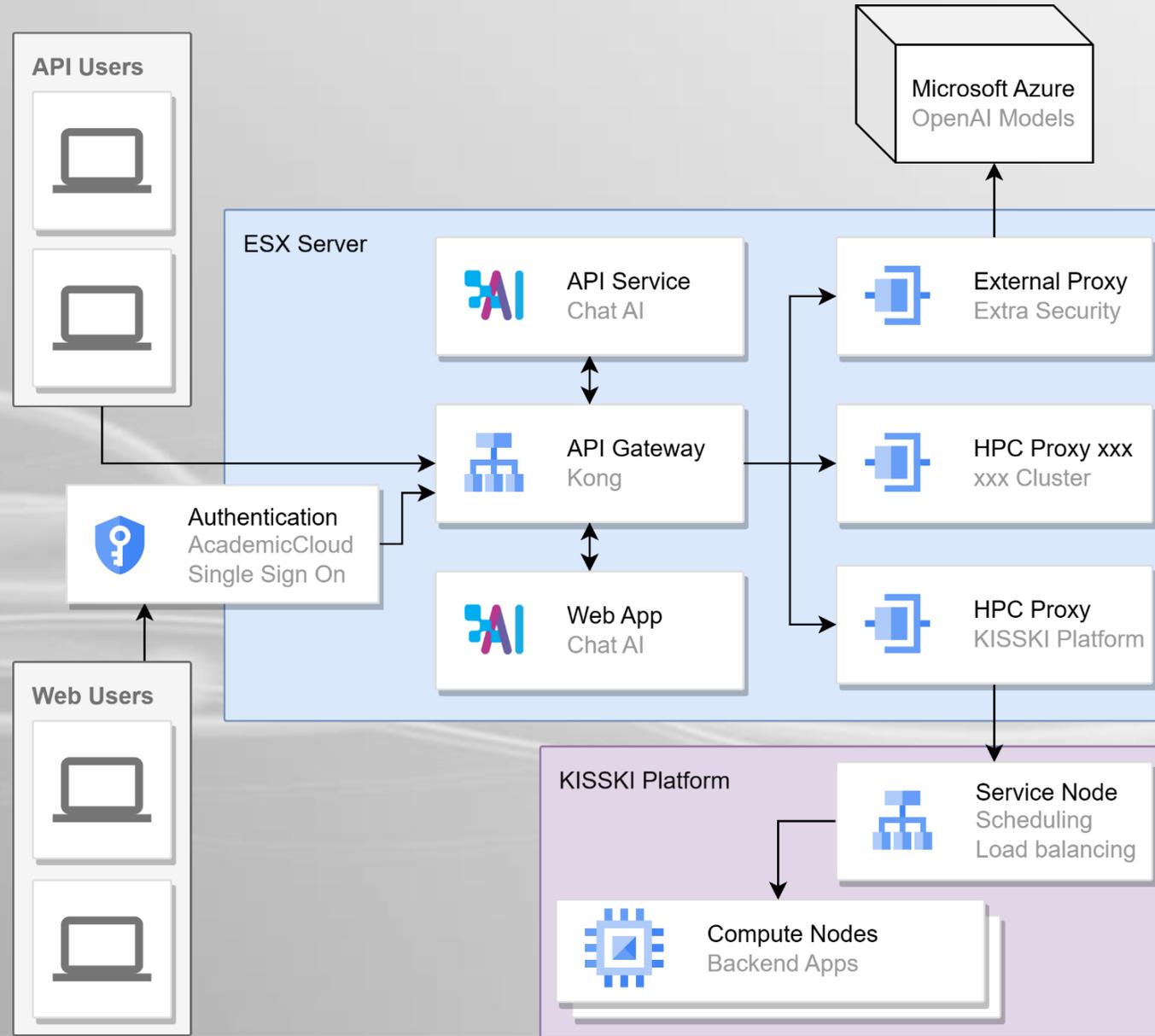
Eigenes Chat AI Frontend



The screenshot displays a web-based chat interface with the following components:

- Header:** Includes a moon icon, the "Chat AI" logo, and navigation links for "KISSKI", "GWDG", and "RY".
- Left Sidebar:** Contains a button for "Add New Chat" and a list item for "Untitled Conversation".
- Warning Box:** A blue-bordered box at the top of the chat area contains the text: "Hinweis: Chatbots sind anfällig für Halluzination und ihre Antworten sollten nicht als korrekt angesehen werden. [Datenschutz](#) und [Impressum](#)." with a close icon.
- Chat Area:** A large white space for the conversation, currently empty, with an upward arrow icon at the bottom right.
- Input Field:** A text input at the bottom with the placeholder text "Fragen Sie mich" and icons for adding attachments and voice recording.
- Settings Panel (Right):** A sidebar for configuring the chat model, including:
 - Modell:** A dropdown menu currently set to "Teuken 7B Instruct R..."
 - Arcana:** Two input fields labeled "id" and "key".
 - temp:** A slider ranging from "Logical" to "Creative".
 - top_p:** A slider ranging from "Focused" to "Diverse".
 - System prompt:** A text area containing "You are a helpful assistant".
 - Buttons:** "Optionen ausblenden", "Teilen" (green), and "Standard zurücksetzen" (grey).

Eigenes Chat AI Backend

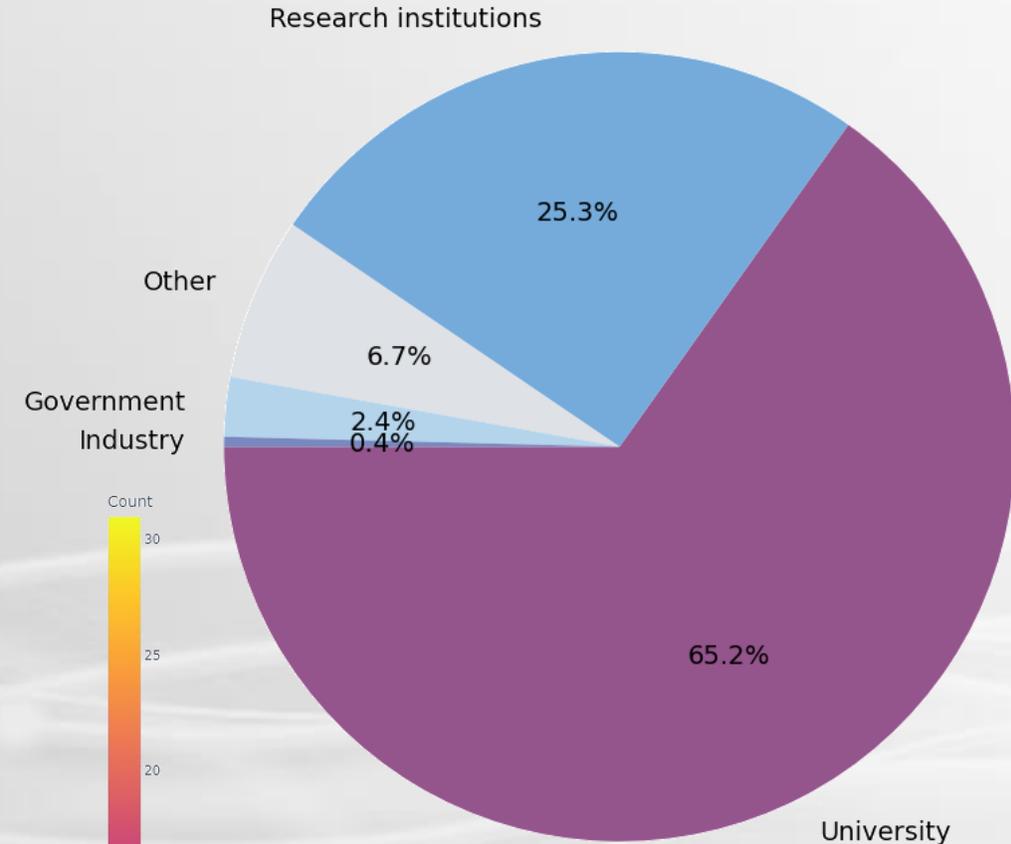
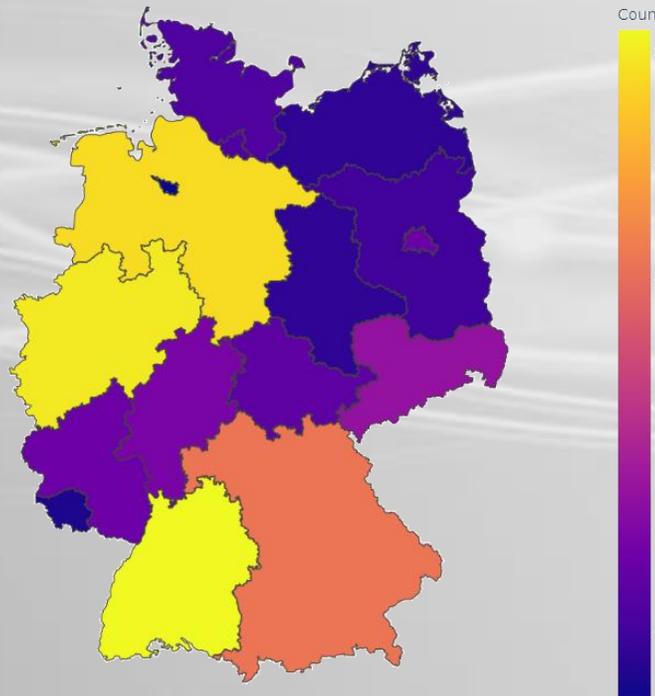


Nutzer von ChatAI

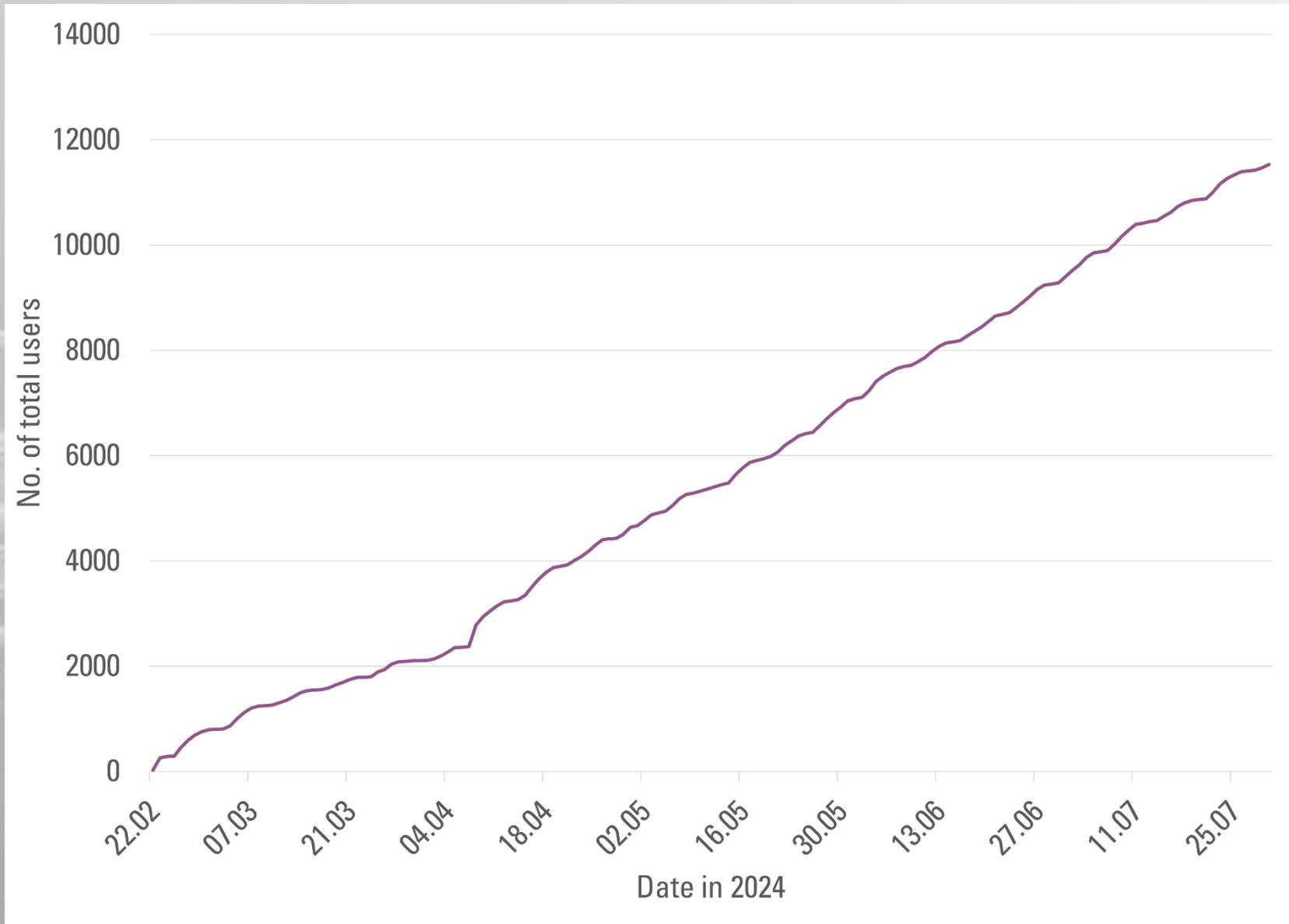
- Verfügbar für Dritte nach Abschluss Vereinbarung
- Aktuell über >180 Universitäten
- Zahlreiche nicht-universitäre Nutzer

Beispiele:

- KI-Lab des Umweltbundesamtes
- Deutsches Krebsforschungszentrum
- Financial Intelligence Unit des Zolls
- FloodWaive (Vorhersage von Überschwemmungsereignissen)
- Gesundheitsforen Leipzig (Akteure digitales Gesundheitswesen)
- Heise Verlag
- Nds Landesinstitut für schulische Qualitätsentwicklung

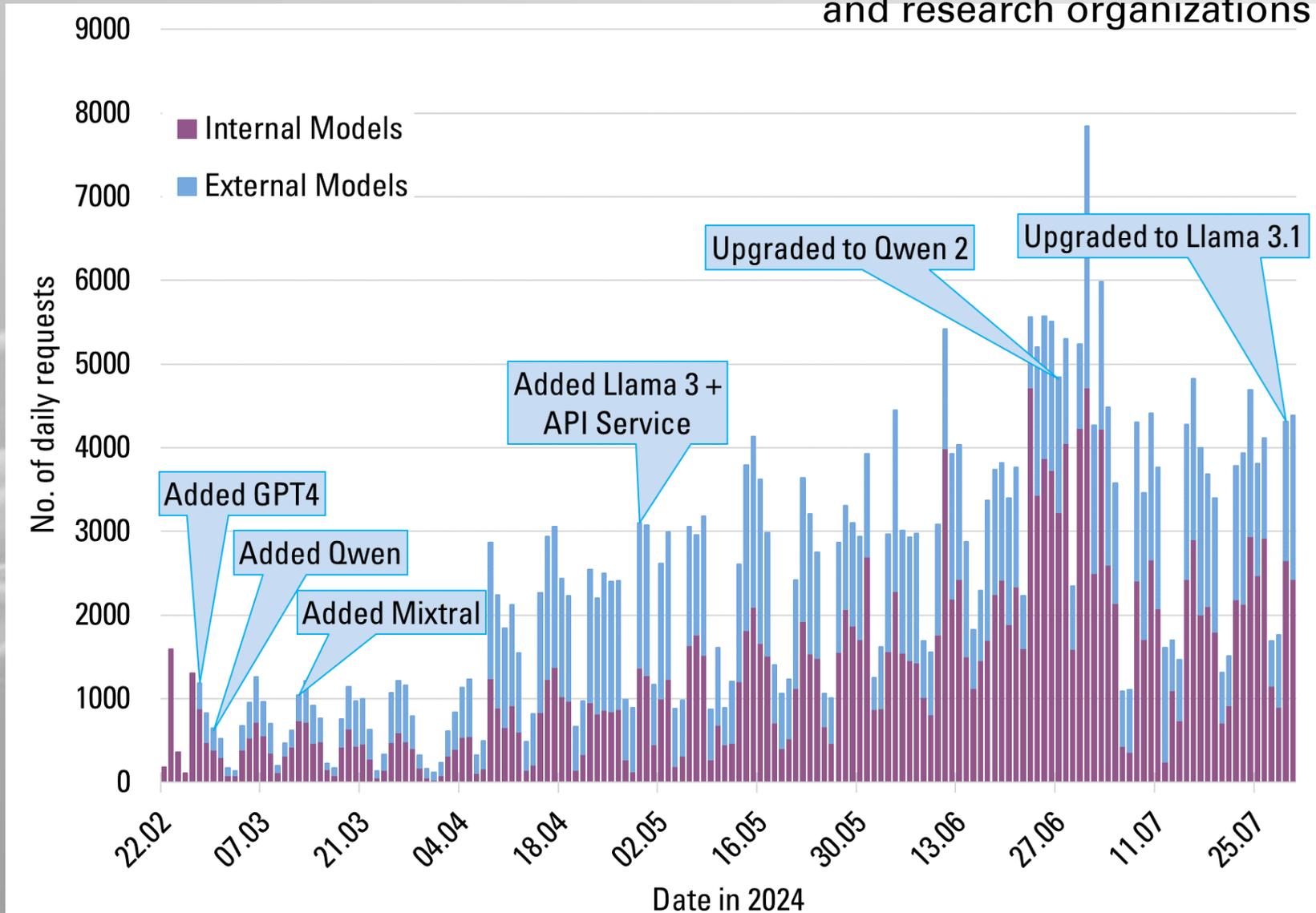


Gesamtzahl individueller Nutzer



Tägliche Nutzer

Used by over 180 universities and research organizations



- meta-llama-3.1-8b-instruct
- meta-llama-3.1-70B-instruct
- qwen2-72b-instruct
- mixtral-8x7b-instruct
- codestral-22b
- e5-mistral-7b-instruct
- llama-3.1-sauerkrautlm-70b-instruct
- occiglot-7b-eu5-instruct
- umg-virtual-assistant (llama 2)
- touken-7b-instruct

Beitrag zu offenem Ökosystem

- Freier Zugriff auf offene Modelle für alle
 - Über kostenlosen AcademicID Account
- API Zugriff möglich
 - Kostenlos via <https://kisski.gwdg.de/> buchen
 - Lediglich Backend AV-Vertrag nötig
 - Nutzbar in Applikationen wie SillyTavern, HAWKI (siehe Talk)
- Weiterhin: Zugang zu ChatGPT4
 - kostenlos für Nutzer in Niedersachsen und aus der MPG
 - Dritte: Kostenübernahme notwendig; Lizenzvertrag für ChatGPT4 über die GWWDG möglich

Ressourcen

- Infrastruktur:
 - LLMs mit 70 Billionen Parametern benötigen je 4 NVIDIA H100 GPUs
 - Anzahl Instanzen skalieren automatisch mit Anfragevolumen
 - Start eines Modells benötigt bis zu 10 min
 - Daher 1 permanente Instanz pro Modell erforderlich für akzeptable Antwortzeit
- ChatAI Web Interface auf der GWDG on-premise Cloud
- 5 Mitarbeiter an Entwicklung, Betrieb und Support
- Sonst als Erweiterung der NHR-Supportstruktur

ChatAI Entwicklungen

Bedarf an eigenen Anpassungen:

- Fine-Tuning Service (LoRA, PeFT, ZenML)
 - Schlüsselfertige Beispiele für die Anpassung der Modelle
 - Management via Webseite (im Aufbau)
 - Beispiel umgesetzte Projekte: UMG Virtual Assistant

- RAG-Systeme zur Kontextanreicherung und Faktensicherstellung
 - Verwalten von Daten in verschlüsselter Form
 - Anwendungsbeispiel: Upload von Vorlesungsmaterialien
 - Ziel: ChatBot für Veranstaltung als Mentor für Studierende
 - Beispiel umgesetzte Projekte: KITA-Hub, Hogrefe Verlag

ChatAI Ökosystem

Weitere Funktionen

- Einbettung in andere Webseiten
 - konfiguriertem Prompt, Einstellungen nach Wunsch und RAG
- Voice AI (Whisper Modell und BBB Integration)
- Coco (Code completion in VS Code via API)
- Image AI (Bildgenerierung)

Ausblick:

- Arbeiten auf C5-Zertifikat
 - aktuell ISO 9001 und 27001

Neu: 08.12.2024

EU-Kommission etabliert sieben AI Factories:

- GWDG beteiligt an dem deutschen Knoten: HammerHAI
- Gemeinsam mit HLRS, LRZ, KIT, SICOS BW
- Ca. 50 Mio. Euro für AI Infrastruktur
- Schnittstelle zu den bestehenden KI Servicezentren
- Beratung und Support
- Unterstützung auch für KMUs

PRESSEMITTEILUNG: 69/2024 | 10.12.2024

Özdemir: AI Factory kommt nach Deutschland – Großer Fortschritt für KI-Infrastruktur in Deutschland und Europa

Deutschland errichtet eine AI Factory am Höchstleistungsrechenzentrum Stuttgart, eines der drei Bundeshöchstleistungsrechenzentren des Gauss Centre for Supercomputing

Das EuroHPC Joint Undertaking (EuroHPC JU) hat heute die Gründung von sieben neuen „AI Factories“ in Europa bekannt gegeben. Am Standort Stuttgart wird eine neue, für KI-Anwendungen optimierte Supercomputing-Infrastruktur bereitgestellt, die um Serviceleistungen für die Nutzerinnen und Nutzer ergänzt wird. Ziel ist es, den Zugang zu leistungsfähigen KI-Technologien für die Forschung, Start-ups, Mittelstand und Industrie sowie den öffentlichen Sektor deutlich zu verbessern. Der Aufbau von KI-Ökosystemen soll gefördert werden.

Dazu erklärt Bundesforschungsminister Cem Özdemir: „Die erfolgreiche Annahme der Bewerbung des Konsortiums unter Führung des Höchstleistungsrechenzentrum Stuttgart (HLRS) für eine der neuen KI-Fabriken ist ein riesiger Erfolg für den Innovations- und Wirtschaftsstandort Deutschland. Wir benötigen mehr Rechenpower, um Barrieren für den Einsatz künstlicher Intelligenz abzubauen und die Wettbewerbsfähigkeit der europäischen Wissenschaft und Wirtschaft nachhaltig zu stärken. Das neue Angebot wird für KI-Forscherinnen und -forscher, KMUs und Startups ganz neue Möglichkeiten eröffnen, fortgeschrittene KI-Modelle, KI-optimierte Produkte und Geschäftsprozesse zu entwickeln. Das Höchstleistungsrechenzentrum Stuttgart verfügt über jahrzehntelange Erfahrung in der Zusammenarbeit mit der Wirtschaft und ist als Standort der neuen KI-Fabrik hervorragend geeignet. Die Ergänzung der Expertise des HLRS durch die Beiträge seiner Partner wird die Fähigkeiten der AI Factory noch weiter erhöhen.“

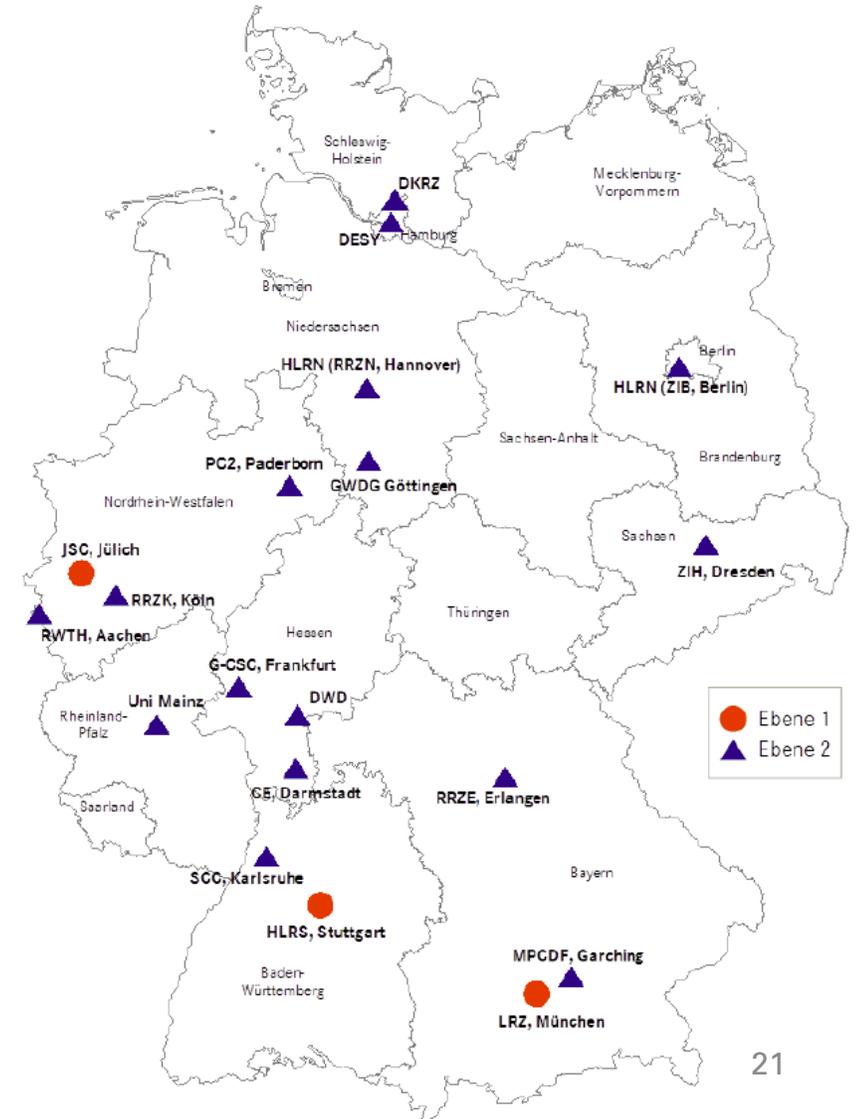
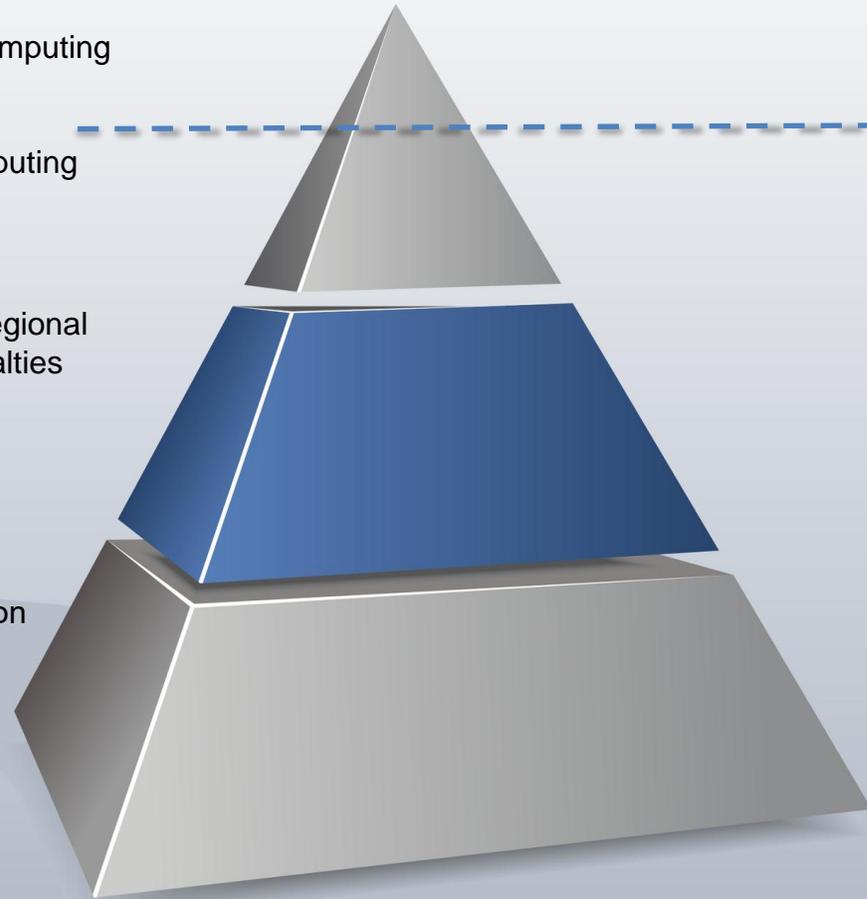
Vergleich zum HPC Ökosystem in Deutschland

Tier 0: European Super Computing

Tier 1: National super computing

Tier 2: HPC centres with regional outreach or thematic specialties

Tier 3: Local HPC centres on university level



Zusammenfassung

- Bedarf an niederschwelliger Zugang zu KI-Lösungen in der Wissenschaft
- Abhängigkeit von kommerziellen Cloud-Services für viele Anwendungen problematisch und hemmend für die Wissenschaft
- KI-Servicezentren bieten einen Ansatzpunkt für eine nationale Infrastruktur
 - Langfristige nationale Strategie zu erwarten
 - Verschränkung mit HPC, Landeslösungen und NFDI, EOSC sind sinnvoll und notwendig
- Zukunftsfelder liegen in der Unterstützung von spezifischen KI-Lösungen (Finetuning, RAG etc.)
 - Beispielsweise in der Lehre, Beratung, Verwaltungsprozessen

- Freier, privater LLM Chat Service mit API
- Viele weitere Services zur Auswahl
- <https://chat-ai.academiccloud.de/>
- Community: <https://gwdg.de./hpc/events/goeaid>



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung