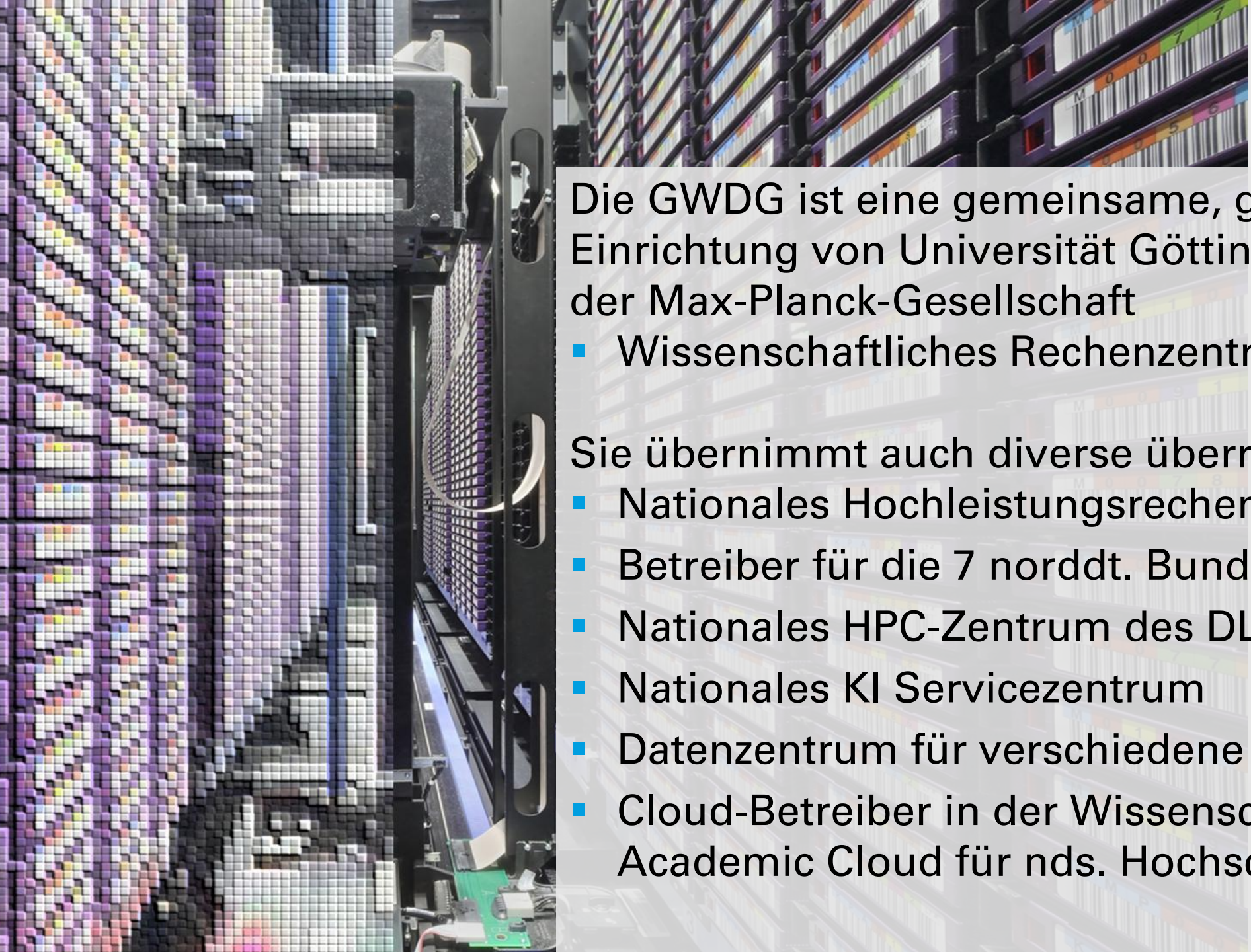
ChatAI - KI-Services in der Wissenschaft

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

Gesellschaft für wissenschaftliche
Datenverarbeitung

Stand 17.12.2024



Die GWDG 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 Daten

KI-Servicezentrum für sensible und kritische Infrastrukturen

GWDG

KISSKI
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

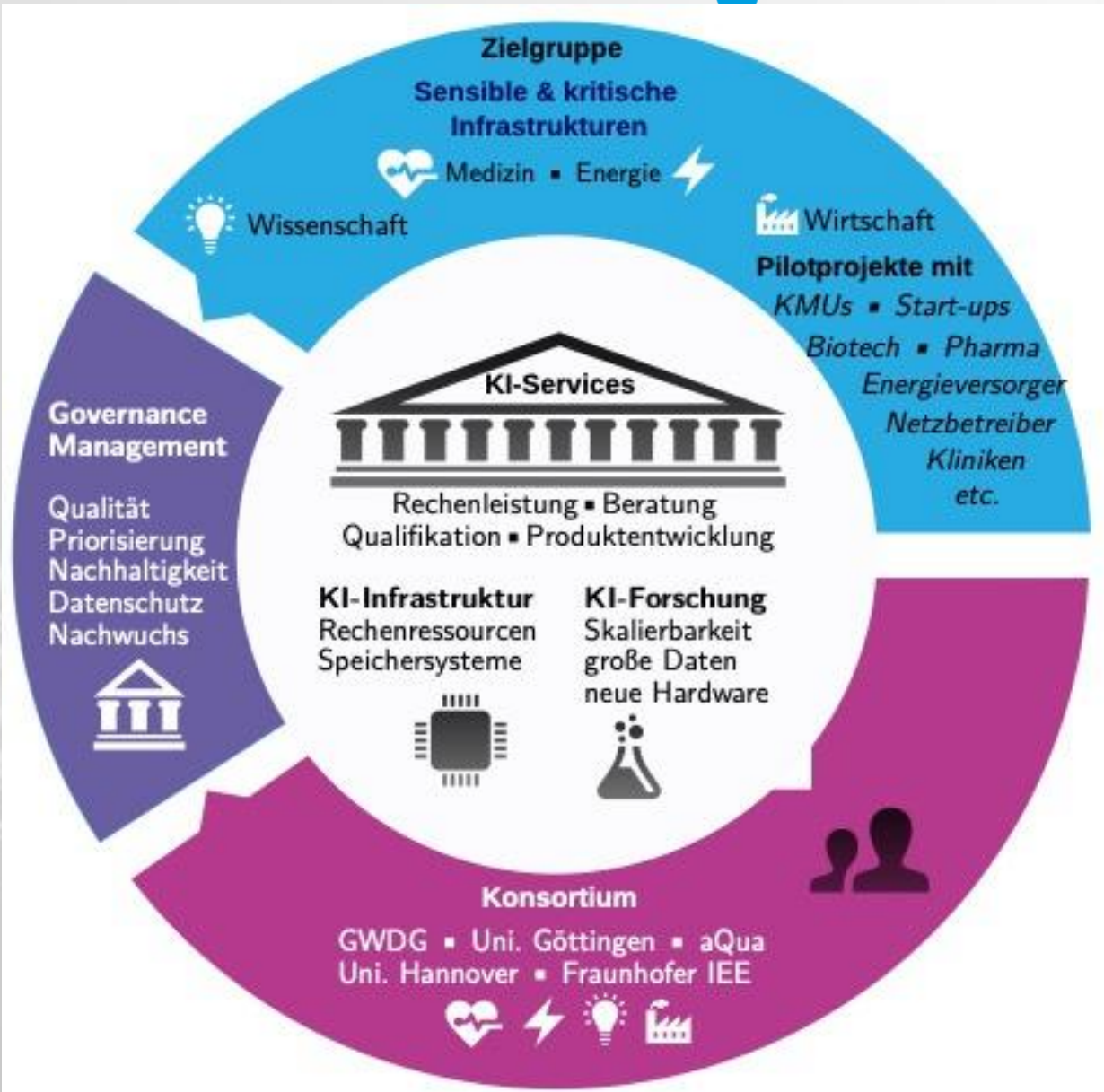
4 Nationale KI Servicezentren

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

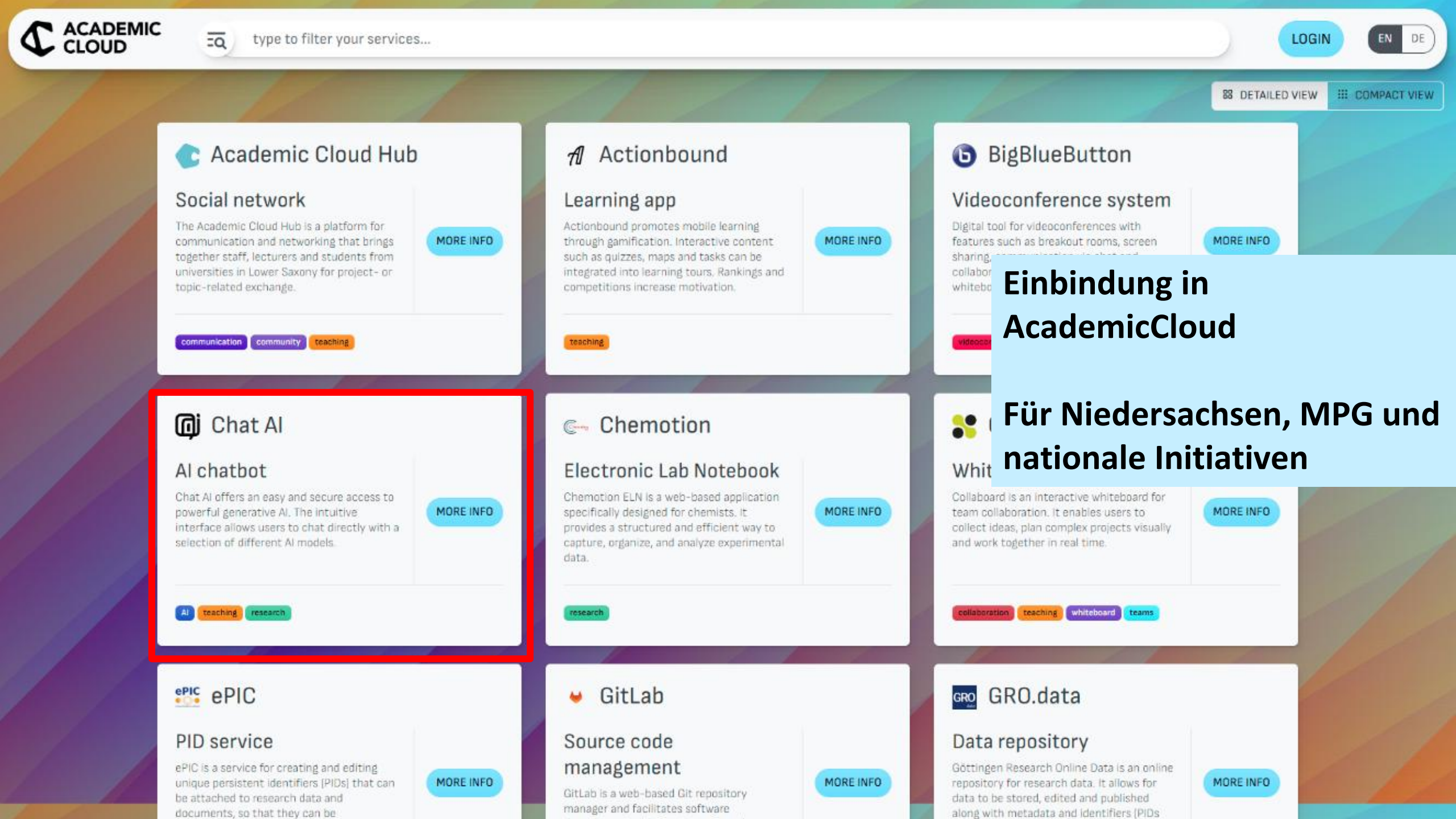
- Bedarf: Forschende & Studierende möchten LLMs nutzen

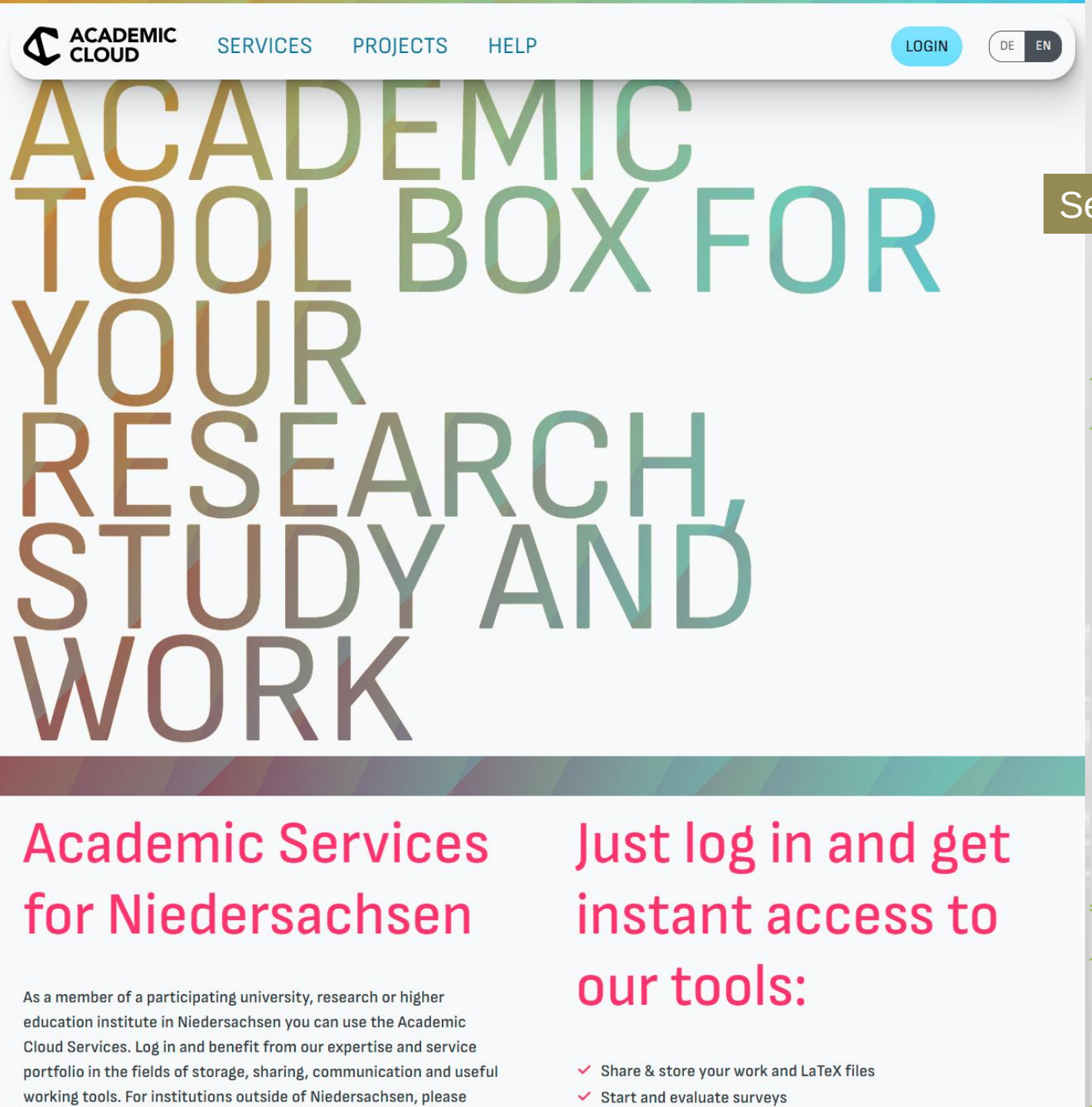
- 1) Angebot an kommerziellen externen Modelle (ChatGPT/OpenAI)
  - Einschränkungen von kommerziellen Anbietern
  - keine Datenschutzgarantie, Unkalkulierbare Kosten mit pay-as-you-go

- 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

https://chat-ai.acedemiccloud.de/


ACADEMIC CLOUD
type to filter your services...
LOGIN
EN
DE
DETAILED VIEW
COMPACT VIEW
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,
MORE INFO
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 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

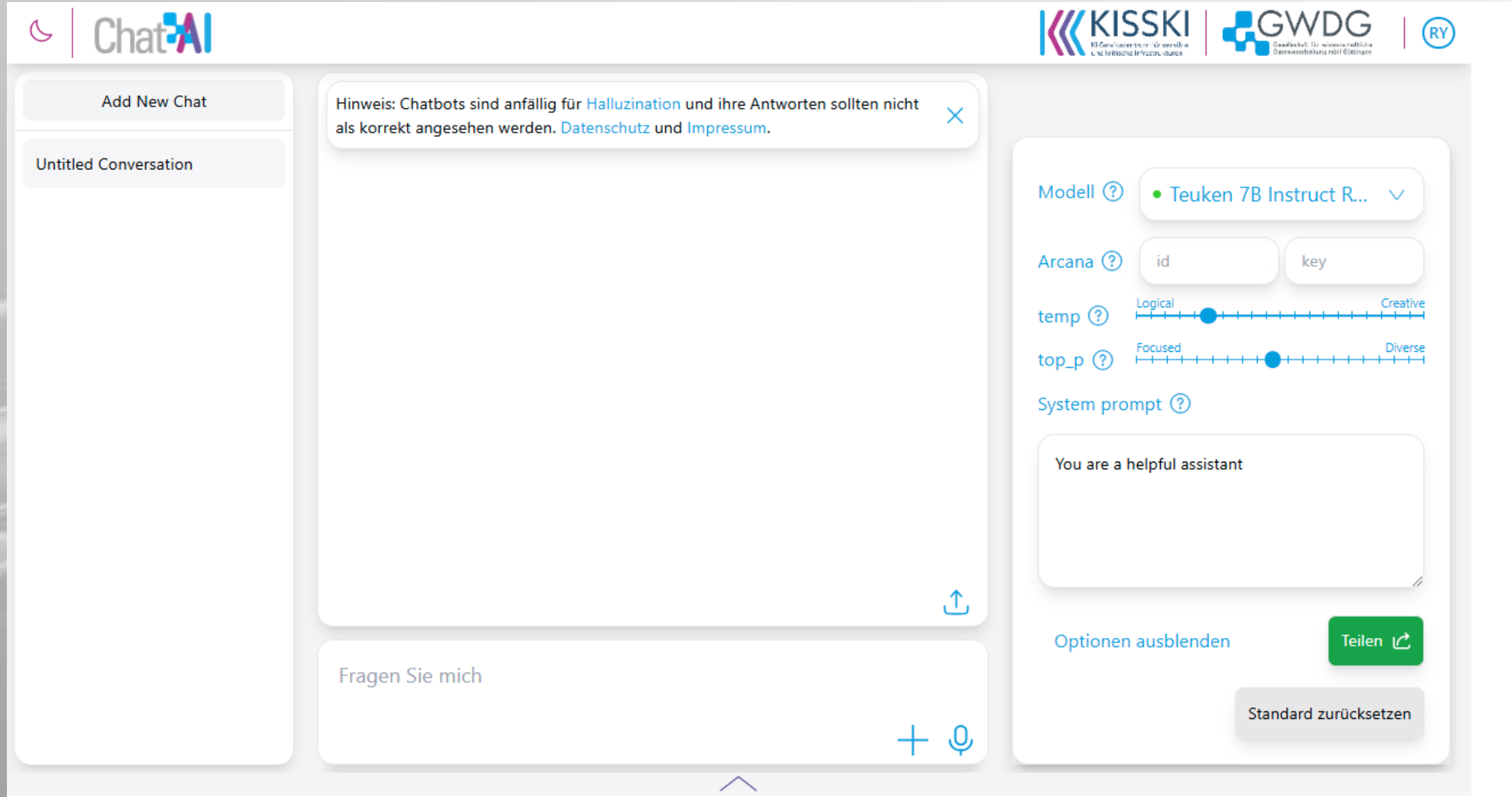


GWDG

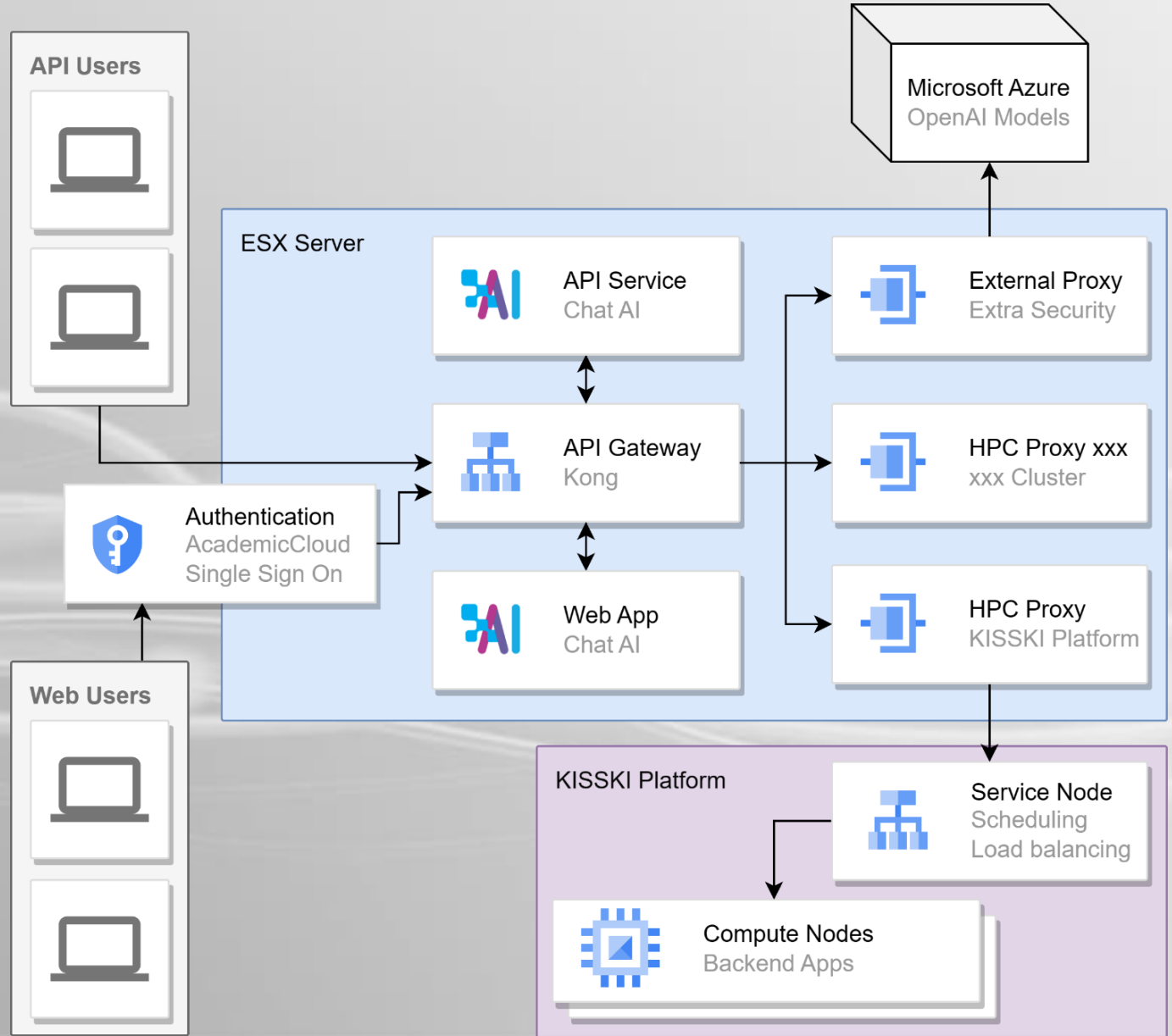
Seit 2015 eine Cloud für die Wissenschaft

Anwendungen	Hintergrunddienste
Sync and Share (ownCloud)	Portal
Gitlab	Dokumentationsserver
ShareLaTeX/Overleaf	IDM + SSO
DataVerse	2FA (Im Testbetrieb)
BigBlueButton	
PID-Service	
Rocket Chat	
OnlyOffice	
JupyterNotebook (Im Testbetrieb)	
Neu	
Elektronisches Laborbuch	**ChatAI**
Knowledge-Graph	Matrix
Collaboard	

Eigenes Chat AI Frontend



Eigenes Chat AI Backend

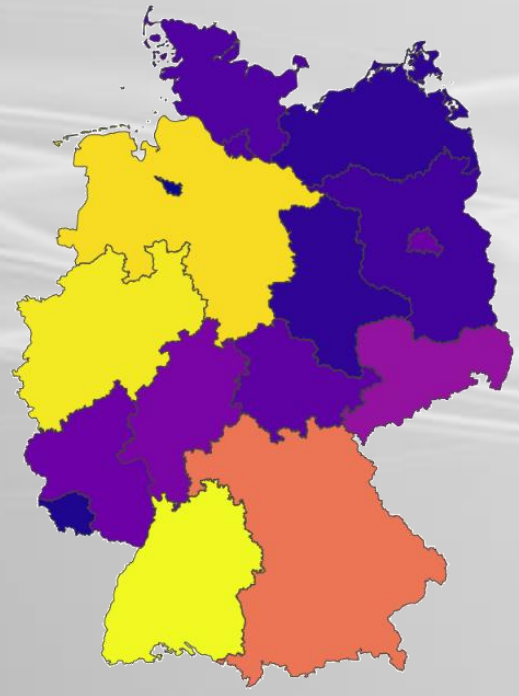


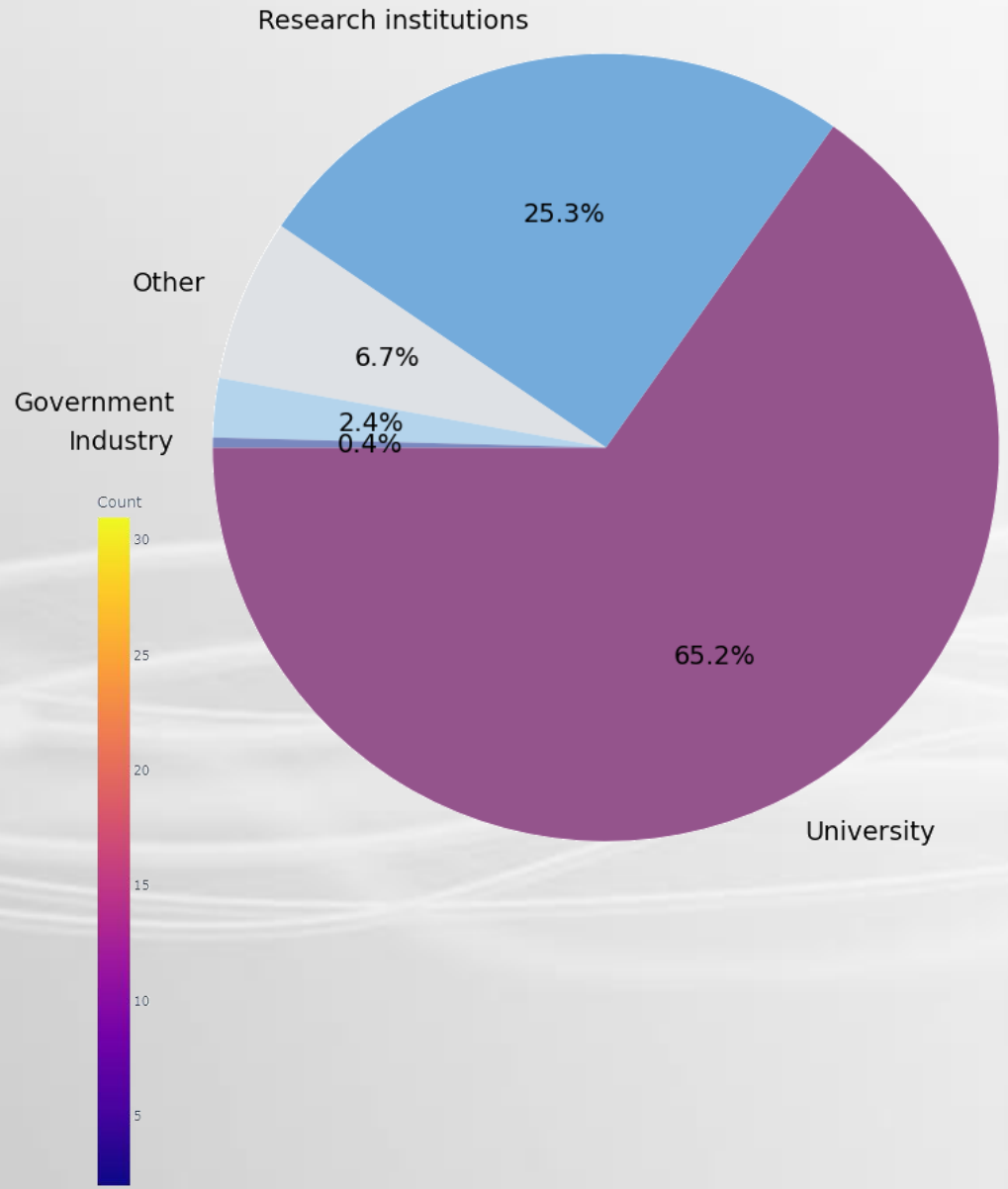
Nutzer von ChatAI

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

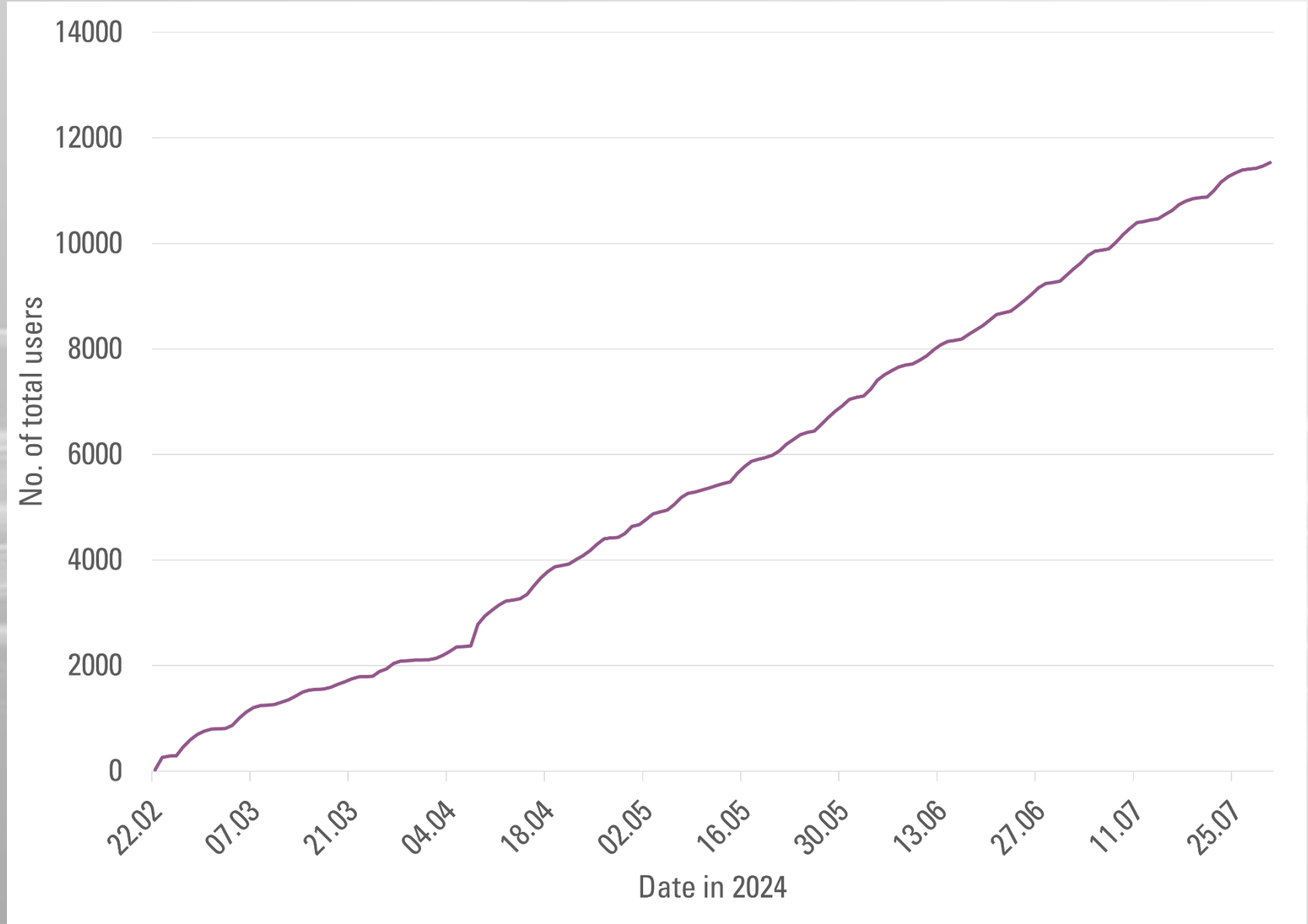
Beispiele:

- 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



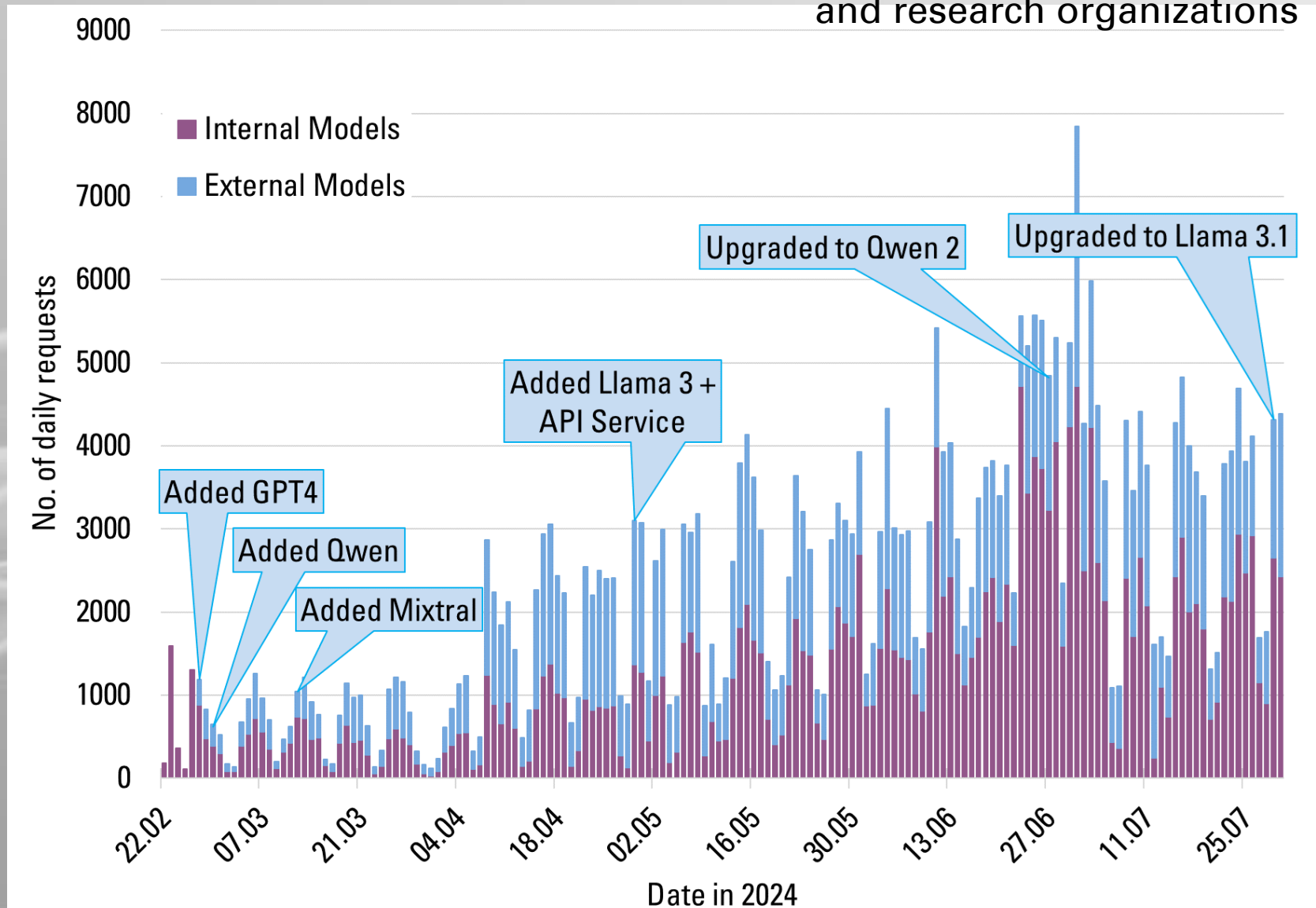


Gesamtzahl individueller Nutzer



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)
- teuken-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 GWDG 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

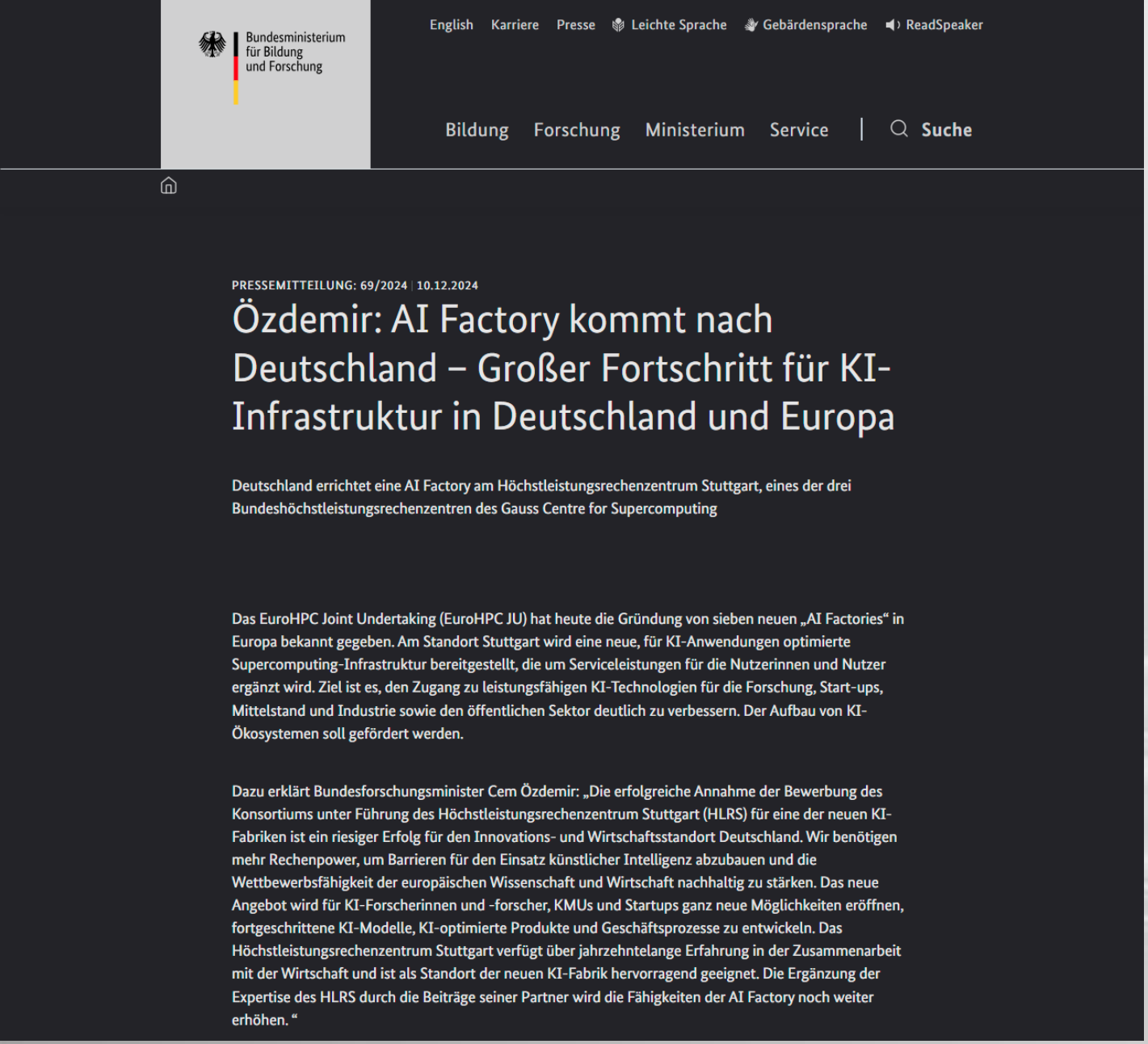
ChatAI Ökosystem

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

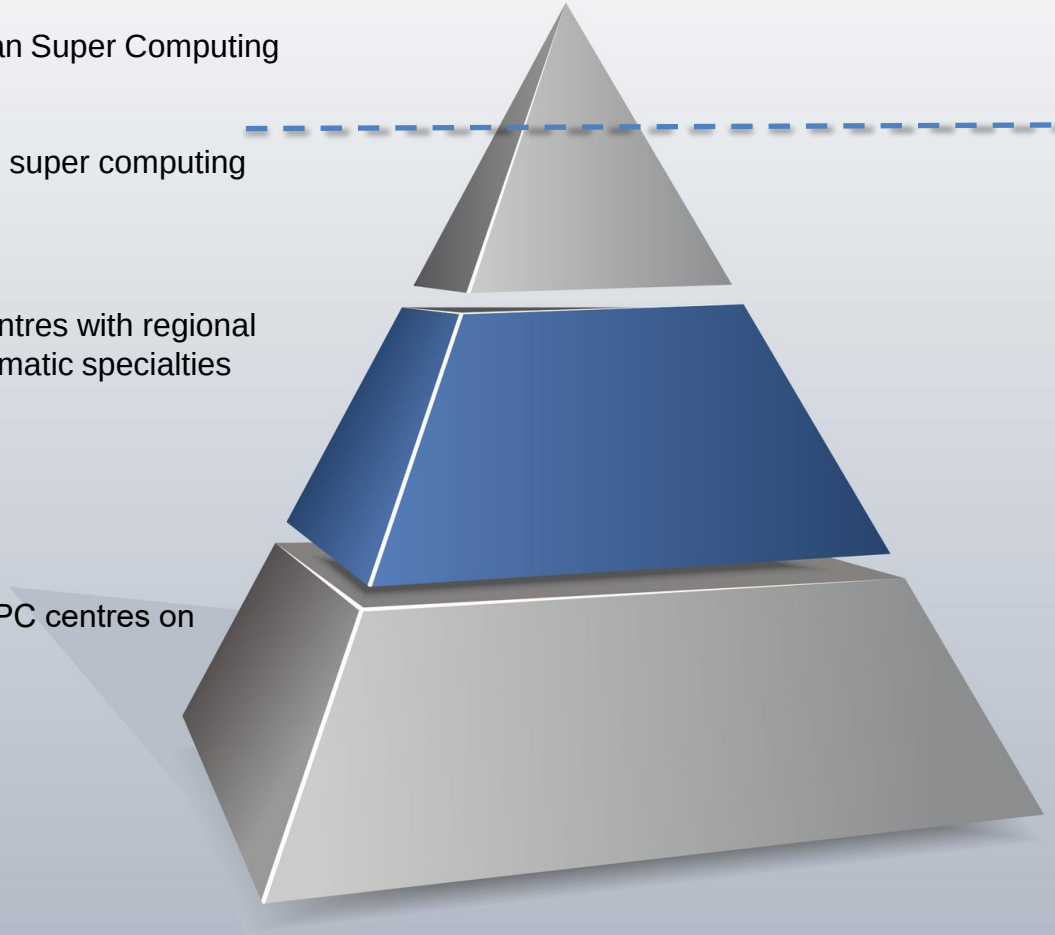
Vergleich zum HPC Ökosystem in Deutschland

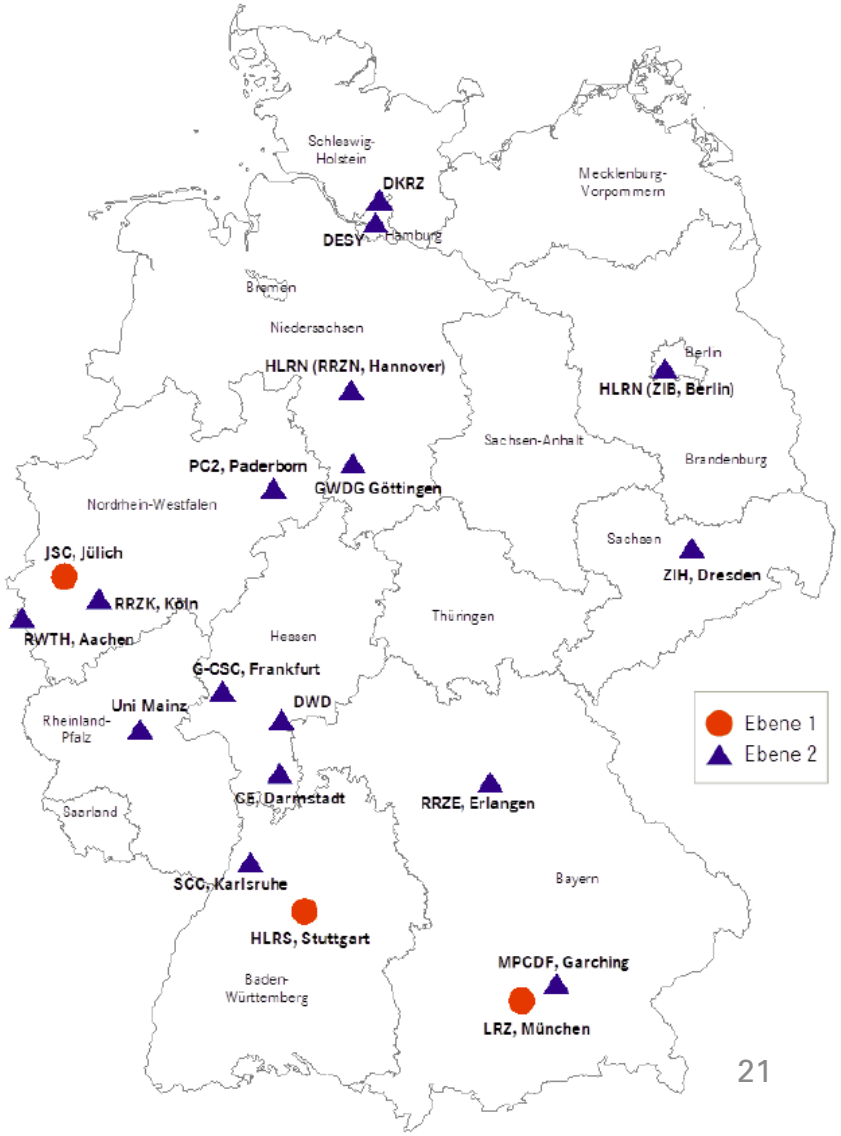
Tier 0: European Super Computing

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