

Christian Boehme

Die KI-Services der GWDG





Services

via



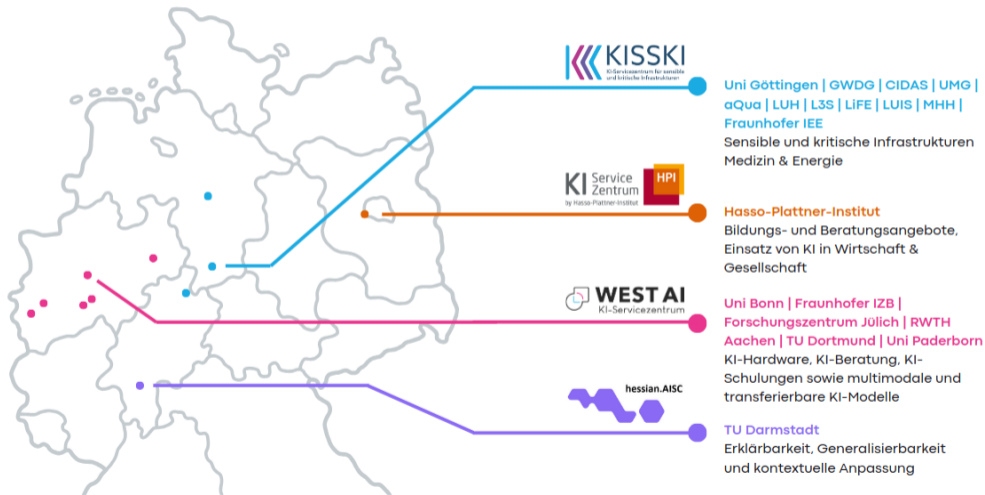
KI-Servicezentrum KISSKI



KISSKI: KI Servicezentrum für Sensible und Kritische Infrastrukturen

- Erfüllung der **Anforderungen** kritischer Infrastrukturen: Sicherheit, Datenschutz, Zuverlässigkeit
- Dienstleistungen für Pilotprojekte in ganz Deutschland
 - ▶ Angebote Rechenressourcen (PaaS) bis Endnutzerdienste
 - ▶ Beratung und Training
- **Forschung** zur weiteren Verbesserung der Dienstleistungen
- Fokus auf die Branchen Medizin und Energiewirtschaft ABER branchenoffen

Eines von 4 nationalen BMBF geförderten KI Servicezentren



Niederschwelliger Zugang - Quick-Start Guide für Neukund:innen

- Einstiegsberatung
- Monatliche KISSKI Online-Einführung
- Direkte Buchung via Webseite
 - 1 Registrierung in der Academic Cloud
 - 2 Auswahl der gewünschten Leistung im [Leistungskatalog](#)
 - 3 Buchung erfolgt auf der Leistungsseite
 - Angabe spezifischer Vorab-Informationen
- Monatliche Community Treffen bspw.
 - ▶ [GöAID](#) - Community für AI Developers
- Handling von Supportanfragen via Tickets

KISSKI
Konsortium für Integrierte Service- und Softwareentwicklung

Über uns | Zielgruppen | Leistungen | Aktuelles

Einstiegsberatung Energie und Gesundheit

Zielgruppen

- Unternehmen (jeder Größe) und Forschungsinstitute aus den Bereichen Medizin und Energie ohne Vorerfahrung zu datengetriebenen Lösungen und Geschäftsmodellen.

Ihre Anforderungen

- Gründliche Erhebung des Use Cases und der Bedarfe
- Aufzeigen erster Möglichkeiten zur Nutzung der vorhandenen Daten
- Vermittlung an passende Fachberater:innen aus dem Konsortium zur weiteren Ausarbeitung

Unser Angebot

Wir bieten eine Einstiegsberatung sowie Unterstützung für Unternehmen und Forschungsinstitute im Bereich Medizin und Energie an, die noch keine praktische Erfahrung mit der Konzeption und Umsetzung datengetriebener Lösungen und Geschäftsmodellen haben. Dabei konzentrieren wir uns auf die Anwendungssysteme Medizin und Energie, und definieren den gewünschten Anwendungsfall so präzise wie möglich im gemeinsamen Dialog. Hierzu diskutieren wir die verfügbaren Datenquellen und darauf aufbauende Modelle vor dem Hintergrund des Use Cases hinsichtlich der Einsetzbarkeit und des Nutzens. Nach erfolgter Beratung und positivem Gutachten werden Services aus der Bereitstellung, dem Consulting oder der Produktentwicklung mit Unterstützung des Servicezentrums beantragt und das Projekt an den passenden Anbieter innerhalb des KISSKI-Konsortiums weitergeleitet.

Nutzungsvoraussetzungen

- Grundsätzliches Verständnis der eigenen Datenstruktur(en)
- (Bestenfalls) Grundlegende Idee über das Zielbild des Anwendungsfalls

Rechtliche Hinweise

Kontakt
Impressum
Datenschutzklärung

Dienstleistungen

Support
FAQ
AGB
Datenschutzklärung zur Auftragsverarbeitung

Weitere Informationen

Pressekontakt
Leistungsangebote
Meine Academic ID
Folgen Sie uns

FAQ | Support | Buchen

Art des Services

Beratung

Ansprechpartner:in

Til Ole Diesterhöft
Felix Kegel

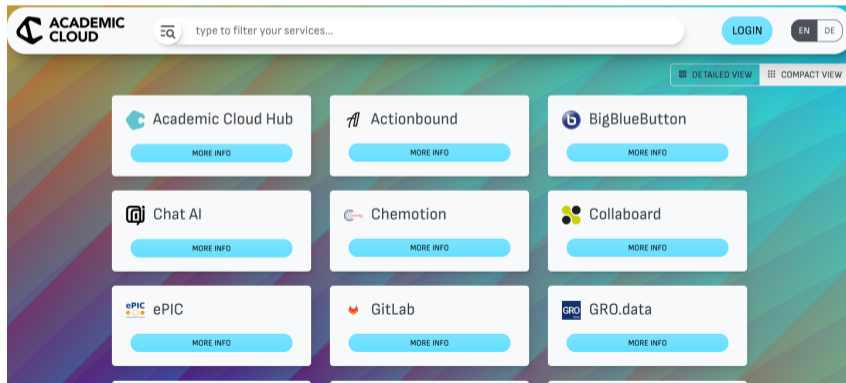
geplanter Starttermin

ab sofort

© 2023 – 2024 OWDO. All rights reserved.

AcademicID

- Universelles Benutzerkonto der Academic Cloud
- Föderierte Anmeldung vieler Universitäten
- Zugang zu geprüften Diensten



Dienste für Endanwender und gehostete Dienste

- (Generative) KI Lösungen nutzbar im Browser
 - ▶ Chat AI, ProteinAI, Voice AI, Image AI, RAG
- API Zugang zu KI Modellen
- AV-Verträge für Chat AI Backend
- Hardwareressourcen

Ziel: Toolkit für Modell-Entwickler zum Webseiten-Hosting für Endnutzende

Chat AI

- **Privat:** Speicherung der Konversation nur im Browser
- **Import/Export** von Konversationen
- **Konfigurierbar:** Systemprompt, temp, top_p
- Einfaches Sharing von guten Einstellungen/Prompts
 - ▶ teilbar als Personas: github.com/gwdg/chat-ai-personas
- Unterstützung von **Vision** Language Models
 - ▶ Nach Auswahl eines VLMs können Bilder hochgeladen werden
- **RAG**-Integration für öffentliche Daten
- Mehrere parallele Gespräche (im Browser gespeichert)
- PDF-Dateien/Bilder können als Kontext mitgeliefert werden

Chat AI - Features

The screenshot displays the Chat AI web interface. The browser address bar shows the URL `chat-ai.academiccloud.de/chat/d047e4b0-1076-4550-8d71-5caf191eea68`. The interface includes a sidebar with 'Add New Chat' and 'Image Analysis' options. The main chat area shows a user query: 'What can you see in the picture?' with an image of a server room. The AI response is: 'In the picture, I can see a cat holding a sign that says "Coming soon! Image AI." The cat is standing on a piece of equipment in what appears to be a data center or server room. The text "GWGD" is displayed prominently at the top of the image.' A second query asks 'What is the color of the text "GWGD"?' and the response is 'The color of the text "GWGD" is light blue.' The right sidebar contains model settings for 'Qwen 2 VL 72B Ins...', Arcana parameters (id, key), temperature and top_p sliders, a system prompt 'You are a helpful assistant', and buttons for 'Optionen ausblenden', 'Teilen', and 'Standard zurücksetzen'.

Chat AI - Features

The screenshot shows a web browser window titled "Chat AI" with the URL `chat-ai.academiccloud.de/chat/9e2bbe9b-03d7-4c23-9975-4e68be97d7aa`. The interface includes logos for KISSKI, GWDG, and HG. A question is asked: "What are data pools at the GWDG?". The answer provides a definition and a list of five key features: shared storage, versioning, metadata, access control, and data documentation. On the right, there are controls for "Anlagen" (885.46 KB), a model selector set to "Meta Llama 3.3 70B In...", and sliders for "temp" (Logical to Creative) and "top_p" (Focused to Diverse). A system prompt area contains the text "You are a helpful assistant".

What are data pools at the GWDG?

Based on the provided documentation, here is a summary of what Data Pools are at the GWDG:

Definition: A Data Pool is a centralized storage space where users can store and manage datasets that are relevant to different user groups over a longer period of time. These datasets can include training data sets for machine learning applications, open data sets from (inter-)governmental organizations, and project data that can be shared with other users.

Key Features:

- 1. Shared storage:** Data Pools provide a shared storage space for users to store their data.
- 2. Versioning:** Each Data Pool has a version, and users can submit new versions or corrections to existing pools.
- 3. Metadata:** Each Data Pool has metadata files (e.g., README.md, METADATA.json, CITATION.bib) that provide information about the dataset, its usage, and licensing.
- 4. Access control:** Access to non-public Data Pools is restricted via ACL (Access Control List), and users can grant access to specific projects or users.
- 5. Data documentation:** Data Pools require documentation, including a

Anlagen Alle löschen

Data Pool __ Documentation for H... 885.46 KB Processed

Modell ? • Meta Llama 3.3 70B In... ▼

Arcana ? id key

temp ? Logical Creative

top_p ? Focused Diverse

System prompt ?

You are a helpful assistant

Hide ↶ zurücksetzen

Komponenten

- Chat AI Webinterface
 - ▶ Eigene Entwicklung basierend auf React und Vitejs
- Middleware
 - ▶ Eigene Entwicklung basierend auf Kong
- HPC scheduler
 - ▶ Eigene Entwicklung basierend auf Slurm
- HPC Rechenressourcen
 - ▶ 21 GPU-Server finanziert via KISSKI
- LLMs
 - ▶ Top open weights Modelle wie Llama3.3, Mistral and Qwen2

Offenes Ökosystem

- **Zugang** über AcademicID Account
- Frontend für Endnutzende, kann auch selbst gehostet werden
- Backend für Entwickler via API Token
 - ▶ **Einfache Buchung** von Token über kisski.gwdg.de
 - ▶ OpenAI-kompatibler **API** Zugriff
- Nutzbar mit **selbstgehosteten** Diensten
 - ▶ ChatAI ist der default Endpoint bei HAWKI
- Bereitstellung von kommerziellen Diensten wie **ChatGPT**
 - ▶ Vertrag mit ChatGPT über uns möglich
- **Open Source** Code + Governance (!)
 - ▶ <https://github.com/gwdg/chat-ai>

RAG Integration - Übersicht

■ **RAG:** Retrieval-Augmented Generation

- ▶ Kombination von Retrieval und Generierung

■ Vektor DB ist integriert in Chat AI

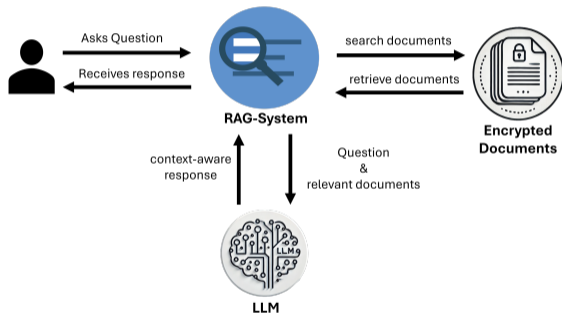
- ▶ Server-Seitig auf GPU-Knoten

■ Vorgehen des RAG-Systems

- ▶ Informationsextraktion aus benutzerdefinierten Dokumenten
- ▶ LLM kann darauf basierend eine Antwort generieren

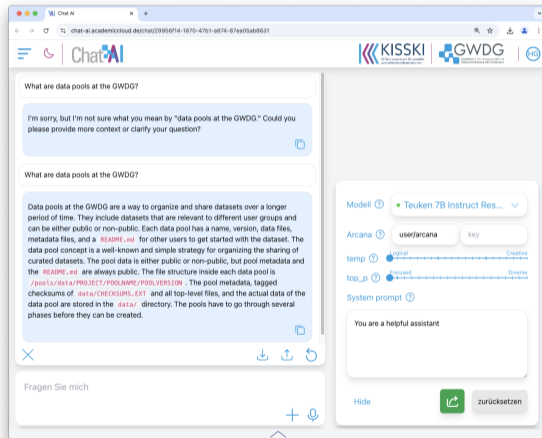
■ Verschlüsselung von Dokumenten

■ Forschung an verbesserter Identifikation relevanter Informationen



RAG Integration: Middleware und Chat Interface

- vLLM nutzt unsere RAG-Middleware
- Lädt bei Anfrage Daten aus verschlüsselter ChromaDB
- Entschlüsselung über "Sharing-Key" des Nutzens
- Prompt wird mit Ausschnitten aus Dokumenten erweitert
- vLLM erzeugt Antwort
- Funktioniert exklusiv mit internen Modellen



RAG Integration: Management Interface

- Erstellung von Dokumentensammlungen, "Arcanas"
- Unterstützung von Text, Markdown und PDF Dateien
- Hinzufügen neuer Daten mittels Embedding Prozess
- Teilen von Arcanas mittels Link (enthält Sharing-Key)

The screenshot displays the Arcana Management web interface. The browser address bar shows the URL `chat-ai.academiccloud.de/arcanas/arcana`. The interface includes a sidebar with 'RAG-Manager' and 'Arcana Management' sections. The main content area features a table of Arcana entries:

ARCANA NAME	EMBEDDED	SIZE (KB)	NUMBER OF FILES	SHARED	STATUS	ACTIONS
GWDG_Docs	Yes	35.14 KB	1	Shared	Ok	Edit Delete Access Link Generate Index
GWDG_SSH	No	6.26 KB	1	Not shared	Currently processing	Edit Delete Access Link

At the bottom left of the interface, there are links for 'Data Privacy', 'Terms of Use', and 'Version 0.2.1'.

Benutzerdefinierte Anwendungen

- OpenAI-kompatibler API Schlüssel kann erstellt werden
- Kann in viele Anwendungen integriert werden
 - ▶ AnythingLLM, RAG auf dem Desktop
 - ▶ LlamaIndex, RAG Bibliothek
 - ▶ DSPY, LLM Programming Frameworking
 - ▶ SillyTavern, Role-play chatting
 - ▶ Und vieles mehr

<https://github.com/Hannibal046/Awesome-LLM>
- Self hosting von eigenen ChatBots, Chat Frontends

Code Completion: CoCo AI

- Code-Vervollständigungsdienst
- Hilft bei Bearbeitung, Erzeugung, Korrektur und Kommentierung von Code
- Integrierbar in **VS Code** und JetBrains mittels des Continue-Plugin
- Zugriff auf alle Chat AI LLMs
- Gleiche Sicherheit und Datenschutz wie Chat AI
- <https://docs.hpc.gwdg.de/services/coco/index.html>



Was kann CoCo AI ?

■ Analysieren

- ▶ Verwendet Snippet/Datei/Codebase als Kontext

■ Generieren

- ▶ Generierung von Code basierend auf Kontext und Aufgabe

■ Reparieren

- ▶ Schlägt Code oder Befehle zur Fehlerbehebung vor

■ Vervollständigen

- ▶ Schlägt Code zur Ergänzung der aktuellen Zeile vor

CoCo AI

The screenshot shows a VS Code editor window with a Python script named `read_trees.py` open. The script imports `sys`, `pandas`, and `json`, and uses `deepcopy` from the `copy` module. It defines a `tree_template` dictionary and a `main` function. The `main` function loads a JSON configuration, reads a CSV file, normalizes coordinates, and writes a new JSON configuration with tree data.

The sidebar on the right shows a list of context items from Codebase, including:

- `prefix_ax(ax)`: This function is used to customize the appearance of a 3D plot. It removes the grid, sets the background color, removes ticks, and adjusts the transparency of the z-axis pane.
- `sample_and_group(npoint, radius, nsample, xyz, points)`: This function is used to sample points from a point cloud and group them based on a specified radius. It returns the sampled points' positions and data.
- `PointNetSetAbstraction` and `PointNetSetAbstractionMsg`: These are classes that inherit from `nn.Module` and are used for point cloud segmentation. They contain methods for initializing the layers and performing forward propagation.
- `Get_Ths(pts_corr, seg, ins, ths, ths_cnt)`: This function is used to calculate thresholds for segmentation.
- `aug_rot(self, xyz, target)`: This function is used to randomly rotate a point cloud around the z-axis.
- `sample_points(self, xyz, target)`: This function is used to sample points from a point cloud. It supports two sampling methods: random and farthest point sampling.
- `main(df, side)`: This function is used to generate a forest map based on the input dataframe and side length.

Image AI

- Text-to-image Generierung
- Erstes Model FLUX.1- schnell^{ab}
- Basiert auf OpenAI-kompatiblen API Server
- Optional: Image-to-image
- Bald nutzbar für Kunden
- <https://image-ai.academiccloud.de/>

^a https://github.com/black-forest-labs/flux/blob/main/model_cards/FLUX.1-schnell.md

^b prompt: "A high performance computing cluster. In the Background a cat sitting on top of the HPC cluster and holding a sign that says 'Coming soon!'. At the top 'GWDG', in clean, simple, light blue letters. In the center of the image 'Image AI' in clean, simple, light blue letters."



Image AI

The screenshot shows the Image AI web interface in a browser window. The browser address bar shows the URL `image-ai.academiccloud.de`. The page title is "Image AI".

Prompt:
A high performance computing cluster. In the Background a cat sitting on top of the HPC cluster and holding a sign that says 'Coming soon!'. At the top 'GWDG', in clean, simple, light blue letters. In the center of the image 'Image AI' in clean, simple, light blue letters.

Advanced Options:
Width: 1280
Height: 720
Models: Will add more models later. Selected: flux
Number of Images: 4

Generated Images:
Four generated images are shown in a row. The first three are thumbnails, and the fourth is a larger preview. The first three thumbnails show a high-performance computing cluster with a cat on top holding a sign that says "Coming soon!". The first two thumbnails also have "GWDG" at the top and "Image AI" in the center. The fourth thumbnail shows a similar scene but with "Image AI" at the top and "GWDG" in the center.

Download full-resolution:
new0.png 839.5 KB ↓ ×
new1.png 878.6 KB ↓ ×
new2.png 832.1 KB ↓ ×
new3.png 922.9 KB ↓ ×

Generate

[Privacy Notice](#) | [Imprint](#)

Voice AI: Transcription and Translation of Uploaded Audio

Transkription & Translation

- Transkribiert Audiodateien in Text
- Erzeugt Videountertitel
- Unterstützt mehrere Sprachen

BBB Integration (In Beta*)

- Ziel: Besprechungen in Echtzeit transkribieren
- Nahtlose Integration in BBB
- Sitzungszusammenfassung
- Erhöht die Inklusivität
Barrierefreiheit

Voice AI

voice-ai.academiccloud.de

voice-ai.academiccloud.de

VoiceAI

KISSKI GWDG

<1500 MB / <60 min audio file to text conversion

Input language

Text format

CHOOSE FILE

TRANSCRIBE IN SOURCE LANGUAGE

TRANSCRIBE AND TRANSLATE TO ENGLISH

Once you submit your job, it will enter the queue. After completion, you can download the results from here.

Input language	Text format	Action	Status	Result
en	text	transcribe	finished	DOWNLOAD
iformiert_08_01_2025_fruh.mp3	de	transcribe	finished	DOWNLOAD

Protein AI

- Ziel: Vorhersage von 3D Strukturen der Proteine
- **Code:** Colabfold
 - ▶ Umfangreiche Datenbanken
 - ▶ Schnelles und empfindliches MMseqs2-Verfahren
 - ▶ Vergleichbare Ergebnisse zu AlphaFold2

The screenshot shows the Protein AI web interface. On the left, there are three 3D protein structure models. Below them is a form with a 'Type' section containing radio buttons for 'Monomer' (selected) and 'Multimer', and a 'Protein Sequence' text area. A blue 'SUBMIT' button is at the bottom of the form. On the right, a table displays search results:

ID	Type	Status	Result	3D Structure
4bdd5247-0e8e-41e3-9f8e-ecff3cb84c2c.fasta	monomer	finished	Download	Show
2ddf1f030-4028-446c-be99-3a48184c1bf3.fasta	monomer	finished	Download	Show

At the bottom of the interface, there are links for 'Privacy Terms', 'Terms of Description', and 'FAQ', along with flags for the UK and Germany.

Protein AI

The screenshot displays the Protein AI web interface with the following components:

- Results Panel:** Shows the protein ID `4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_unrelaxed_rank_001_alphafold2_stm_model_a_seed_000.pdb`.
- Structure Viewer:**
 - Sequence of `https://protein-ai.academiccloud.de/...` with a chain view.
 - Structure Tools: Includes options for Structure (URL, Type, Model), Quick Styles (Default, Stylized, Illustrative), Components (Preset, Add, Polymer, Carbon), and Measurements (Add, Export Animation).
 - 3D visualization of the protein structure.
- Model Analysis Panel:**
 - Structure Analysis:**

with 49 residues, and it is composed of a single chain with no identified ligands. The presence of 3 estimated helices and 2 estimated beta sheets suggests that the protein has a compact, globular fold, which is consistent with many types of proteins.

To determine the specific function of this protein, additional information would be necessary, such as:

 1. Sequence analysis: Amino acid sequence comparison with known proteins could provide clues about its function.
 2. Literature search: Looking up the protein's name or accession number in databases could reveal its known function.
 3. Biochemical assays: Experimental data from assays that measure the protein's activity or interactions could provide direct evidence of its function.

Without this additional information, I can only speculate about the protein's function based on its structural features. If you have more information about the protein, I'd be happy to try to help you further.
 - Input field: "Ask about the structure..."
- Sequence Coverage Plot:**

4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_coverage.png

Heatmap showing sequence coverage (0.0 to 1.0) across 50 positions.
- Pairwise Distance Matrices:**

4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_pae.png

Five heatmaps showing predicted pairwise distance matrices (PDM) for different models.
- Root Mean Square Deviation (RMSD) Plot:**

4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_rmsd.png

Line graph showing predicted RMSD per position (0 to 100) across 50 positions for models `rank_1`, `rank_2`, `rank_3`, and `rank_4`.

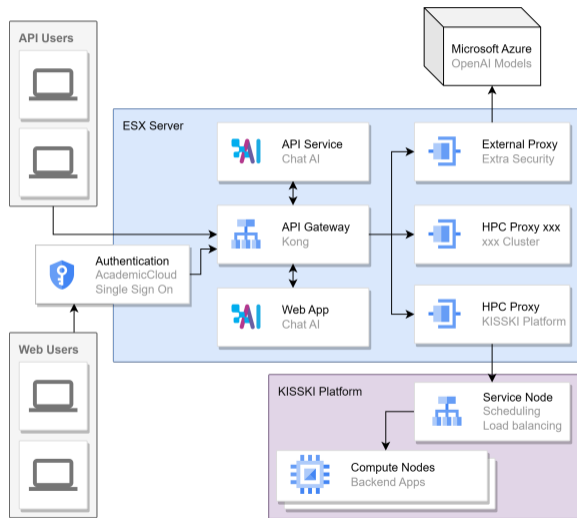
Weiterentwicklung: Scalable AI Accelerator

SAIA Plattform und Ökosystem

- Stand: Die bisherigen Lösungen für Gen-KI-Dienste wachsen zusammen
- Ziel: Entwicklung zur **Plattform** für KI Inferenz Services
 - ▶ Für eigene Services: Chat AI, CoCo AI, Voice AI, Image AI, ...
 - ▶ Nutzende können auch eigene Services hosten
 - ▶ Können über KISSKI oder an anderen HPC Zentren gehostet werden
 - ▶ SDK für Entwicklung an beliebigen Zentren / lokal
- Zusätzliche **Beratung** zur Bereitstellung und Anwendung von KI
 - ▶ Benutzerdefinierte Daten in LLM via RAG
 - ▶ Strategien zur Einhaltung des Datenschutzes (EU AI Act)
- **Informationssicherheit:** ISO 27001 zertifizierte Plattform
- Hosting der Modelle von Trainingsplattform und mehr!

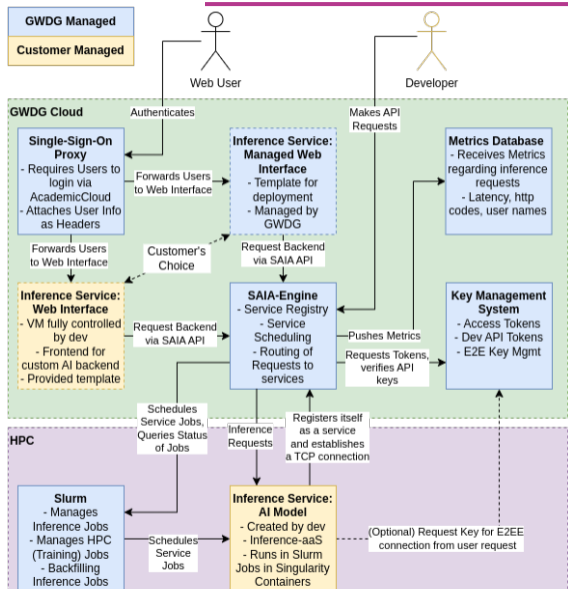
SAIA 0.9 - Aktuelle Architektur - Interne Abläufe

- Proxy übernimmt Authentifizierung
- API Gateway wickelt gesamte interne Routing ab
- HPC Proxy stellt SSH-Tunnel zu HPC
- Verwendet SSH ForceCommand, fungiert als "circuit breaker"
- Bspw. vLLM läuft in Slurm-Jobs



SAIA 2.0 Architecture

- Kundenservice mit Frontend in Cloud und Backend auf HPC
- Web interface auf Cloud VM
- KI Modelle von Kunden in Singularity Containern auf HPC
- Anfragen Ende-zu-Ende Verschlüsselt über KMS
- SAIA als Broker nutzen und Integrität garantieren



SAIA 2.0

- Weiterentwicklung der Lösung in 2025
- **Ziel:** Platform-as-a-Service (PaaS) für Inferenz
- Bereitstellung von KI services auf unserer SAIA Plattform durch Kunden
 - ▶ Benutzerfreundliches Hosting von Inferenz-Modelle
 - ▶ Zusätzlich: Hosting von Webinterfaces/Frontends für Inferenz-Modelle
- Vorteile:
 - ▶ Self-Service
 - ▶ Automatische Skalierung
 - ▶ Automatisches Monitoring
 - ▶ Nutzung von HPC-Hardware
 - ▶ Nutzung von Accounting und User-Management
- API Key self-service, Developer Interface: <https://saia.gwdg.de/>

Ressourcennutzung der Inferenzplattform

- Instanzen für ChatAI skalieren automatisch basierend auf dem Anfragevolumen
 - ▶ Der Start eines Modells dauert bis zu 10 min
 - ▶ Eine aktive Instanz für akzeptable Antwortzeiten erforderlich
 - ▶ Integration in SAIA 2.0 für alle Dienste geplant
- 21 Server mit insgesamt 4 H100 verfügbar
- Webinterface(s) läuft in On-Premise Cloud
- 5 Mitarbeiter kümmern sich um Entwicklung, Betrieb und Support
- 70B LLMs erfordern jeweils 4 NVIDIA H100 GPUs
 - ▶ Mit reduzierter Genauigkeit (Quantisierung): 2

Notwendigkeit

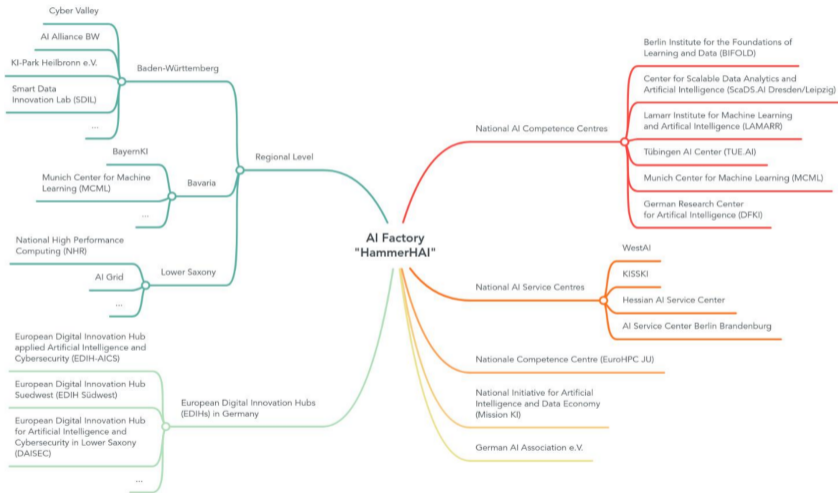
- Datensicherheit - Datenverarbeitung in der EU (vs. USA/China)
- Stärkung des Standorts Europa in Forschung und Industrie
- Wettbewerb im KI Sektor muss JETZT stattfinden
- EU-weite Strategien verringern Rechtsunsicherheiten
 - ▶ Auch ausländische Unternehmen müssen EU-Standards einhalten
 - ▶ KI-Fabriken, KI-Gigafabriken

GWDG im KI-Ökosystem

NHR-NORD@GÖTTINGEN

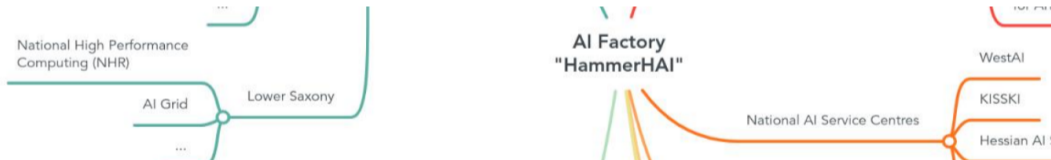


HammerHAI



HammerHAI

- Start der EU KI Fabrik "HammerHAI" 2025
- Ergänzung zum nationalen KI-Servicezentrum KISSKI
- Stärkung des Standort Göttingen und der Region Südniedersachsen



Outline

1 KISSKI

2 ChatAI

3 CoCo AI

4 Image AI

5 Voice AI

6 Protein AI

7 SAIA

8 EU KI-Ökosystem

9 Rechtliches

10 Zusammenfassung

Rechtliches

Relevant für Nutzende

- Urheberrecht
- DSGVO
- KI-VO
- Institutionen: Pflichten von Anbietern

Relevant für Anbieter

- Urheberrecht
- DSGVO
- KI-VO
- Pflichten von Betreibern
- Subventionsrecht

Urheberrecht

- In EU und USA KI-generierte Inhalte oft nicht urheberrechtlich geschützt, da kein menschlicher Urheber existiert
- Nur Werke mit menschlicher Kreativität genießen urheberrechtlichen Schutz
- KI-Training auf urheberrechtlich geschützten Daten in der EU ist nur unter bestimmten Voraussetzungen möglich (z.B. nicht-kommerzielle Forschung)
- Nutzende werden häufig als verantwortliche Partei betrachtet, insbesondere wenn der Output geschütztes Material enthält
- Das Problem ist nicht KI-spezifisch sondern analog zu externen Quellen

“Die **Nutzer:Innen** des Service “Chat AI” haben die **vollständige und alleinige Verantwortung** die geltenden Bestimmungen des Urheberrechts zu beachten und einzuhalten. Die GWDG weist die Nutzer:Innen explizit darauf hin, dass die bereitgestellten **Sprachmodelle von Dritten trainiert** wurden und der GWDG keine Erklärung vorliegt, die die verwendeten Materialien auf freie Lizenzen einschränkt. Es kann folglich von der GWDG **nicht ausgeschlossen** werden, dass die bereitgestellten Sprachmodelle mithilfe urheberrechtlich geschützter Inhalte trainiert wurden. Antworten, die die Sprachmodelle den Nutzer:Innen geben, **können folglich urheberrechtlich geschützte Inhalte beinhalten**. Die GWDG weißt explizit die Nutzenden darauf hin, dass ein **direktes Weiterverwenden von erhaltenen Antworten nicht empfohlen** ist. Die **Prüfung des Urheberrechts** für solche Fälle liegt **alleine bei den Nutzer:Innen**. Die GWDG übernimmt keinerlei Haftung für etwaige Schadensersatzansprüche aus Urheberrechtsverletzungen.”

KI-VO

- Regulierung von KI zur Sicherstellung von Sicherheit, Transparenz und Ethik
- Anwendungsbereich: Entwicklung, Bereitstellung und Nutzung von KI
- Betrifft Betreibende, Anbietende und Nutzende von KI
- Strafe: Bis zu 6 % des globalen Jahresumsatzes

Regulierung nach KI-VO

- KI-Systeme mit minimalen oder keinem Risiko
 - ▶ KI in Videospiele, Spamfilter
- KI-Systeme mit begrenztem Risiko
 - ▶ Chatbots, Empfehlungssysteme
- Hochrisiko-KI-Systeme
 - ▶ KI in kritischen Infrastrukturen, Personalmanagement, Kreditwürdigkeitsprüfung, Diagnosen von Patienten
- Verbotene KI-Systeme (unannehmbares Risiko)
 - ▶ Social Scoring, Profiling zur Vorhersage von Straftaten, Unterschwellige Beeinflussung von Personen, Ausnutzung der Schutzbedürftigkeit von Personen, Gesichtsdatenbanken, Emotionserkennung am Arbeitsplatz/Bildungseinrichtung (außer aus medizinischen oder Sicherheitsgründen)

Handlungsempfehlungen für Nutzende

- Datenschutz und rechtliche Fragen klären
 - ▶ Verwendung von KI-Systemen Dritter nur bei gültiger Rechtsgrundlage
- KI-Output immer kritisch hinterfragen (z.B. Halluzination)
- KI-Anwendungen nur aus seriösen Quellen installieren
- Links und Anhänge, die eine KI ausgibt, prüfen
- Human-in-the-Loop-Prozesse integrieren

Outline

1 KISSKI

2 ChatAI

3 CoCo AI

4 Image AI

5 Voice AI

6 Protein AI

7 SAIA

8 EU KI-Ökosystem

9 Rechtliches

10 Zusammenfassung

Ausblick - insbesondere für ChatAI

- Verbesserungen des RAG-Systems
- VoiceAI Steuerung
- Tool Integration Berechnung auf dem HPC System
- Hosting von Fine-Tuned Modellen (LoRA)
- ChatAI Train - Extra Werkzeug zur Sammlung von Trainingsdaten/Feedback für LoRA und RAG
- Semi-Structured Interviewer - Agentensystem für die Integration
- OmnisHPC - Steuerung von HPC Experimenten und tiefe HPC Integration

Die Community und Zusammenarbeit ist essentiell

Zusammenfassung

- KISSKI entwickelt ein KI-Servicezentrum für kritische Infrastrukturen
 - ▶ Fokus auf Sicherheit, Privatsphäre, Zuverlässigkeit
 - ▶ Wettbewerbsfähige KI-Infrastruktur Wissenschaft & Industrie
- Die Dienste sind einfach über den Dienstkatalog nutzbar
- F&E Aktivitäten unterstützen Etablierung innovativer Dienste
- KISSKI leistet einen unschätzbaren Beitrag für die nachhaltige Entwicklung von KI Diensten in Deutschland
- Viele Weiterentwicklungen sind geplant

Kosten für die OpenAI Nutzung

- Zugang zu ChatGPT4 ist kostenlos für
 - ▶ Nutzende in Niedersachsen
 - ▶ MPG Angehörige
- Verträge können über uns abgeschlossen werden
- Bisher wurde über 21.000€ an Azure gezahlt



Nutzungsverträge für AI-Service

Chat AI Basis

interne (Open-Weight) Modelle

- ausgefüllter und unterschriebener Auftragsverarbeitungsvertrag
- ausgefüllte und unterschriebene Leistungsvereinbarung
- ausgefülltes Formular zur Benennung des Datenschutzbeauftragten
- Alle notwendigen Dokumente zum Download: <https://owncloud.gwdg.de/index.php/s/PcC7DdkYtfYSzSm>

Chat AI Reseller

externe (kommerzielle) Modelle

- alle erforderlichen Dokumente wie bei ChatAI Basis
- ausgefülltes und unterschriebenes Formular zur Ernennung von IDM-Fachverantwortlichen
- Alle notwendigen Dokumente zum Download: <https://owncloud.gwdg.de/index.php/s/DWLh7qxknKg1EWY>

Preismodell OpenAI Nutzung

Pauschaler Mindestbetrag

- Wir geben die Kosten aus Azure zuzüglich einer Verwaltungspauschale weiter. Hierfür bieten wir einen Pauschalbetrag ab 250,- € netto monatlich an.

Deckelbetrag

- Auf Wunsch kann einen Deckelbetrag vereinbart werden. Sobald der Deckelbetrag erreicht ist, schränken wir den Zugang ein. Die Nutzung der Open Source-Modelle ist davon nicht betroffen.

Tatsächliche Nutzung

- Wir berechnen den Aufwand anhand der tatsächlichen Nutzung (Tokens) der verschiedenen OpenAI Modelle zu den jeweils aktuellen Preisen im Monat der Nutzung.

Mehraufwand

- Wenn die tatsächliche Nutzung den Pauschalbetrag übersteigt, stellen wir den Mehraufwand zusätzlich in Rechnung.