

# Einfluss der KI auf die Lehre der Zukunft

Julian Kunkel, Henrik Brosenne, Jonathan Decker,  
Matthias Eulert



# Agenda

- Willkommen
- Workshop Organization inkl. kleine Survey
- KI-Angebote in Göttingen
- Brainstorming von KI in der Lehre
  - ▶ Prozesse und Akteure
  - ▶ Chancen + Herausforderungen
- Zukunftsvisionen
- Was sollten wir tun?

# Motivation

- KI
  - ▶ Werkzeug – Drastische Produktivitätssteigerung bei richtiger Nutzung
  - ▶ Persönlicher Assistent und Unterstützung, Barrierefreiheit
  - ▶ Bietet Chancen und birgt Risiken
  - ▶ Wird genauso wenig wie Google wieder verschwinden...
- (Generative) KI hat signifikanten Einfluss auf Lehre und Gesellschaft  
Wir/Universität wird sich innerhalb der nächsten 10 Jahre wandeln
  - ▶ Wie wird das Studium der Zukunft aussehen?
  - ▶ Welche Rolle haben Lehrende? Wer ist noch betroffen?
  - ▶ Was bringen wir Studierenden bei?

# Interaktivität

- Workshop Charakter: Interaktion, jede/r darf sprechen
- Vollständigkeit ist nicht so wichtig
- Wir machen Notizen, arbeiten kollaborativ im Pad  
<https://pad.gwdg.de/geister>  
Bitte öffnen
- Miteinander sprechen - Kleingruppen

# Brainstorming

- Kleingruppen, d.h. 3-5 Personen sammeln zuerst (einfach umdrehen)  
Stellt euch auch kurz vor!
- Sammlung der Ergebnisse gemeinsam im Plenum
- Diskussion
- Sofern möglich, gerne in Pad direkt notieren

# Table of contents

- 1 Willkommen
- 2 ChatAI
- 3 Other Services
- 4 Custom Services
- 5 Community

# GWDG and KISSKI

## ■ GWDG

- ▶ IT-service provider for Uni Gö and MPG
- ▶ Providing national storage and compute resources
- ▶ E.g. “Emmy” system:
  - Top500 06/2024: **#156**
  - Green500 06/2024: **#258**



## ■ KISSKI

- ▶ 1 of 4 AI service centers in Germany
- ▶ BMBF funded since 10/2022
- ▶ Various services via [kisski.gwdg.de](https://kisski.gwdg.de)



# Motivation

We provide an LLM-Service with

- Scalable infrastructure
- Open source base
- Built-in data privacy  
and information security (ISO-certified)



Result: ChatAI

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



# Access to the AI-Service-Center

The screenshot displays the KISSKI AI-Service-Center website. The browser address bar shows the URL `kisski.gwdg.de`. The website header includes the KISSKI logo and navigation links for "About us", "Target Groups", "Services", and "News". Language selection buttons for "DE" and "EN" are also present.

The main content is organized into two sections: "Hardware" and "Software".

**Hardware Section:**

- Computing Resources - Training Platform:** GPU-based HPC system with current NVIDIA A100 and H100 GPUs for training tasks.
- Computing Resources - Inference Platform:** GPU-based HPC system with current NVIDIA H100 GPUs and a software environment for inference tasks.
- Computing Resources - Future Technology Platform:** Architectures such as ARM and RISC-V and other heterogeneous hardware systems such as Intel Graphcore.
- Secure HPC Partition:** Isolated partition for processing highly sensitive data (e.g. health data) on our systems, e.g. our GPU-based HPC system with current NVIDIA A100 and H100 GPUs.

**Software Section:**

- Chat AI:** AI chat service similar to ChatGPT with several available models, including ChatGPT-4, and high data protection.
- Secure Container Registry:** Container Registry for the secure HPC partition, which is for instance comprised of GPU-based HPC systems with current NVIDIA A100 and H100 GPUs for training.
- Protein structure prediction:** Ready-to-use software stack and community support for protein structure prediction.
- Voice AI:** Voice AI service offering advanced transcription and translation capabilities, including real-time captioning with high data protection.

# Access to the AI-Service-Center

The screenshot shows a web browser window displaying the KISSKI website. The browser's address bar shows the URL 'kisski.gwdg.de'. The website header includes the KISSKI logo and navigation links for 'About us', 'Target Groups', 'Services', and 'News'. Language selection buttons for 'DE' and 'EN' are visible. The main content area is titled 'Chat AI' and is divided into several sections: 'Target group' (Everyone), 'Your requirements' (a list of conditions for free use), 'Our offer' (a paragraph describing the service), 'Service type' (Software), 'Contact person' (Ali Doost Hosseini and Jonathan Decker), and 'Planned start date' (from now on). A sidebar on the right contains icons for 'FAQ', 'Support', and 'Book'.

**KISSKI**  
KI Servicezentrum für sensible  
und kritische Infrastrukturen

About us Target Groups Services News DE EN

## Chat AI

### Target group

Everyone

### Your requirements

- AI chat without saving your chat history
- Free use of OpenAI GPT-4 (for members of MPI and Lower Saxony)
- Free use of various open source models
- Managed hosting of your own language models
- Fine-tuning of LLMs on your data
- Retrieval Augmented Generation (RAG) based on your knowledge base

### Our offer

We offer you the free use of various Large Language Models (LLM) in a simple chat interface. For users in Lower Saxony or members of Max Planck Institutes, the free use of OpenAI GPT-3.5 and OpenAI GPT-4 is also possible. When using our internal models, your conversations are never stored on the server side. With the OpenAI models, Microsoft may store your conversations for up to 30 days to prevent abuse. However, your data is never

FAQ Support Book

### Service type

Software

### Contact person

Ali Doost Hosseini  
Jonathan Decker

### Planned start date

from now on



Note: The chatbot is prone to [hallucination](#) and its responses should not be considered accurate. [Data Privacy and Imprint](#). ✕

What is Hello World in Pytorch?

In PyTorch, "Hello World" is a simple neural network that prints "Hello, World!" to the console. It's a traditional introductory example in the world of deep learning, and PyTorch is no exception.

Here's the code:

```
import torch
import torch.nn as nn

class HelloWorldNet(nn.Module):
    def __init__(self):
        super(HelloWorldNet, self).__init__()
        self.fc1 = nn.Linear(1, 128) # input layer (1) -> hidden layer (128)
        self.fc2 = nn.Linear(128, 1) # hidden layer (128) -> output layer (1)

    def forward(self, x):
        x = torch.relu(self.fc1(x)) # activation function for hidden layer
        x = self.fc2(x)
        return x

model = HelloWorldNet()

input_tensor = torch.tensor([[1.0]]) # input tensor
output = model(input_tensor)
print(output)
```

Let's break it down:

1. We create a neural network class that inherits from `nn.Module`.



Ask me



Model ⓘ

Meta LLaMA 3 70B Instruct



Advanced options ⓘ

# Features

- Built-in data privacy and information security
  - ▶ User requests and responses are never saved on the server
- Powerful web chat interface
  - ▶ Speech recognition
  - ▶ Up- and download of conversations
  - ▶ Selection of LLMs via drop-down
  - ▶ System prompt configurable
  - ▶ Integrated access to ChatGPT4
- API access via standard API (OpenAI-compatible)
- Selection of open source models

# Contribution to an Open Ecosystem

- Free access to open models for all
  - ▶ Requires an AcademicCloud account
- API access possible
  - ▶ Freely bookable via [kisski.gwdg.de/](https://kisski.gwdg.de/)
  - ▶ Usable in external applications, e.g., SillyTavern, HAWKI
- ChatGPT4 access is free for users from Lower Saxony and the MPG
  - ▶ Contract for ChatGPT possible through us

# Outline

- 1 Willkommen
- 2 ChatAI
- 3 Other Services**
- 4 Custom Services
- 5 Community

# Scalable AI Accelerator - SAIA Platform and Ecosystem

- Code completion
- Speech recognition
- Image generation
- ChatAI extensions in progress
  - ▶ Support for Vision Language Models (VLM)
  - ▶ Integration of RAG for public data
- Consulting on the application of AI
  - ▶ Custom data in LLM via RAG
  - ▶ Data privacy compliant strategies
- Fine-Tuning
  - ▶ 35 Server with 4 A100 80 GB GPUs via KISSKI available
- Certified ISO 27001

# Code Completion: CoCo AI

- A code completion service provided by GWDG through the KISSKI project.
- Assist with editing, generating, fixing and commenting code.
- Operates in Visual Studio Code via the Continue plugin.
- Accesses the same LLMs that ChatAI has access to.
- Same security and data privacy that ChatAI provides.





# What can CoCo AI do?

## ■ **Analyse**

Use snippet/file/codebase as context to ask question.

## ■ **Generate**

Generate in-line code for context-snippet with requested task.

## ■ **Fix**

Suggest solution code or commands to fix errors.

## ■ **Autocomplete**

Suggest code based on file content.

# VoiceAI: Transcription and Translation of Uploaded Audio

- Transcribes audio files to text
- Generates video captions
- Supports multiple languages

The screenshot displays the VoiceAI web interface. At the top, there is a navigation bar with the VoiceAI logo, a moon icon, and logos for KISSKI and GWDG. Below the navigation bar, there are two tabs: "AUDIODATEI TRANSKRIPTION" (selected) and "STREAMING AUDIO TRANSKRIPTION". The main content area is titled "Audio Dateien in Text umwandeln" and contains two dropdown menus: "Sprache" (set to English) and "Textformat" (set to Normaler Text). Below these are three buttons: "DATEI AUFWÄHLEN" (with a file icon), "TRANSKRIPTION IN QUELSPRACHE", and "TRANSKRIBIEREN UND ÜBSETZTEN INS ENGLISCHE". A red "ABORT" button is also present. On the right side, there is a "Ergebnis" section showing the transcription result: "Now go away or I shall taunt you a second time!". At the bottom of the interface, there are links for "Datenschutz", "Nutzungsbedingungen", "FAQ", and "Kontakt", along with a German flag icon and the copyright notice "© 2024 GWDG | Copyright". A "HERUNTERLADEN ALS TXT" button is located at the bottom right of the result area.

# VoiceAI: Live BigBlueButton (BBB) Captioning

- Meetings transcribed real-time
- Meeting summary
- Enhances inclusivity
- Seamless integration in BBB

The screenshot shows the VoiceAI web interface. At the top, there is a navigation bar with the VoiceAI logo on the left, a moon icon in the center, and logos for KISSKI and GWDG on the right. Below the navigation bar, there are two tabs: 'AUDIODATEI TRANSKRIPTION' and 'STREAMING AUDIO TRANSKRIPTION', with the latter being selected. The main content area is titled 'BBB live Transkription'. Below the title, there is a short paragraph explaining the service: 'Dieser Dienst transkribiert den Ton im BBB Raum nach der Benachrichtigung und stellt das Ergebnis in die angegebene Pad-URL. Nach Beendigung des Dienstes erhalten Sie die Zusammenfassung der Sitzung sowohl im Pad als auch in der Zusammenfassungsbox.' Below this text are three input fields: 'Raumadresse @', 'Zugriffsschlüssel @', and 'Korrekturen @'. At the bottom of the input fields are two blue buttons: 'TRANSKRIPTOR STARTEN' and 'TRANSKRIPTOR STOPPEN'. To the right of the input fields is a box titled 'Terminzusammenfassung @' with the text 'Noch keine Zusammenfassung verfügbar.' At the bottom of the page, there are links for 'Datenschutz', 'Nutzungsbedingungen', 'FAQ', and 'Kontakt', along with a copyright notice '© 2024 GWDG | Copyright' and a small flag icon.

# Protein AI

## ■ Code: Colabfold

- ▶ **Rich databases**
- ▶ **Fast and sensitive MMseqs2 method**
- ▶ **Comparative results to Alphafold2 (Full system)**

The screenshot shows the Protein AI web interface. At the top, there are logos for Protein AI, KISSKI, and GWGD. Below the logos, there are tabs for "STRUCTURE PREDICTION" and "DEV". The main content area is titled "Type" and has two radio buttons: "Monomer" (selected) and "Multimer". Below this is a text input field labeled "Protein Sequence". At the bottom, there is a table with a blue header "SUBMIT".

ID	Type	Status	Result	3D Structure
a68aeca9-3bf4-41ad-8990-aefe3fe1bd72.txt	monomer	in queue	in queue	in queue
164c8310-0804-46ef-9c30-2c5ac2de9d7e.txt	monomer	in queue	in queue	in queue
db7e4670-e21f-42cd-90df-d84efc189fbc.fasta	monomer	finished	<a href="#">Download</a>	<a href="#">Show</a>
26a811cc-a22d-41e6-955c-b0dcb4be38a5.fasta	monomer	in queue	in queue	in queue
b4c27a21-a4e2-433d-801f-7a30f2ecb0cd.fasta	monomer	finished	<a href="#">Download</a>	<a href="#">Show</a>

# Image AI

- Text zu Bild Generierung
- erstes Modell FLUX.1 - schnell<sup>ab</sup>
- OpenAI-kompatibler API-Server auf der KISSKI-Inferenzplattform
- (optionale Funktion) Bild-zu-Bild

<sup>a</sup> [https://github.com/black-forest-labs/flux/blob/main/model\\_cards/FLUX.1-schnell.md](https://github.com/black-forest-labs/flux/blob/main/model_cards/FLUX.1-schnell.md)

<sup>b</sup> 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

Not Secure 0.0.0.0:7860

## Image AI GWDG

Prompt

Researchers of all genders and backgrounds get a presentation on a new service for image generation with generative AI. On the slides a logo for the image generator service is shown with "GWDG" is written in clean, light blue letters in the center.

Advanced Options

Width 1280

Height 720

Models  
Will add more models later

flux.1-schnell

Generate

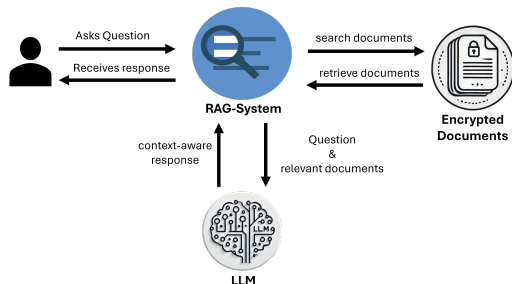
Download full resolution

new0.png	1.1 MB
new1.png	836.1 KB
new2.png	1.1 MB
new3.png	1.1 MB

Use via API - Built with Gradio

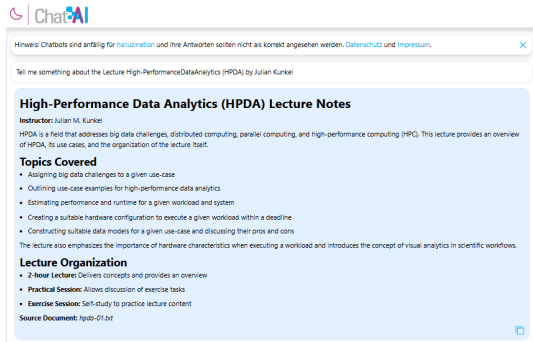
# RAG Integration: Overview

- Our RAG Integration combines information retrieval with language generation for more accurate and context-aware responses.
- Retrieval System
  - ▶ Retrieves encrypted documents and embeddings.
- Language Model (LLM)
  - ▶ Receives the retrieved data and generates responses tailored to user queries based on the relevant data.



# RAG Integration: Middleware and Chat Interface

- VLLM uses our RAG Middleware to connect to the relevant ChromaDB.
- It retrieves embeddings and documents
- The data is then sent to the LLM to generate responses.
- Responses are delivered back to users through our Chat-AI web interface.



The screenshot shows the Chat-AI web interface. At the top left is the Chat-AI logo. Below it is a warning message: "Hinweis: Chatbots sind anfällig für Halluzination und ihre Antworten sollten nicht als korrekt angesehen werden. Datenschutz und Impressum." with a close button. The user's query is "Tell me something about the Lecture High-PerformanceDataAnalytics (HPDA) by Julian Kunkel". The response is a light blue box containing the following text:

**High-Performance Data Analytics (HPDA) Lecture Notes**  
**Instructor:** Julian M. Kunkel  
HPDA is a field that addresses big data challenges, distributed computing, parallel computing, and high-performance computing (HPC). This lecture provides an overview of HPDA, its use cases, and the organization of the lecture itself.

**Topics Covered**

- Assigning big data challenges to a given use-case
- Outlining use-case examples for high-performance data analytics
- Estimating performance and runtime for a given workload and system
- Creating a suitable hardware configuration to execute a given workload within a deadline
- Constructing suitable data models for a given use-case and discussing their pros and cons

The lecture also emphasizes the importance of hardware characteristics when executing a workload and introduces the concept of visual analytics in scientific workflows.

**Lecture Organization**

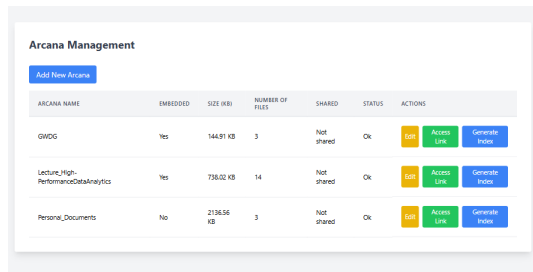
- **2-hour Lecture:** Delivers concepts and provides an overview
- **Practical Session:** Allows discussion of exercise tasks
- **Exercise Session:** Self-study to practice lecture content

**Source Document:** hpdg-01.txt



# RAG Integration: Management Interface

- The system allows users to register.
- Users can create multiple arcanas.
- Files can be imported into the system.
- The system generates encrypted embeddings for each arcana.
- Users can securely share arcanas through a link that includes a decryption key for access.



**Arcana Management**

[Add New Arcana](#)

ARCANA NAME	EMBEDDED	SIZE (KB)	NUMBER OF FILES	SHARED	STATUS	ACTIONS
GWDG	Yes	144.91 KB	3	Not shared	Ok	<a href="#">Edit</a> <a href="#">Access Link</a> <a href="#">Generate Index</a>
Lecture_High-PerformanceDataAnalytics	Yes	730.02 KB	14	Not shared	Ok	<a href="#">Edit</a> <a href="#">Access Link</a> <a href="#">Generate Index</a>
Personal_Documents	No	2136.56 KB	3	Not shared	Ok	<a href="#">Edit</a> <a href="#">Access Link</a> <a href="#">Generate Index</a>

# Outline

- 1 Willkommen
- 2 ChatAI
- 3 Other Services
- 4 Custom Services**
- 5 Community

# Custom Applications

- OpenAI-compatible API keys freely available
- Can be integrated into many applications
  - ▶ AnythingLLM, RAG on desktop
  - ▶ LlamaIndex, RAG library
  - ▶ SillyTavern, Role-play chatting
  - ▶ And much more

<https://github.com/Hannibal046/Awesome-LLM>
- **Train and host your own model ...**
  - ▶ ... and make it available in a customized service.

# Fine tuning with ZenML

- Fine tuning adapts pre-trained models to specific tasks or datasets.
- Benefits:
  - ▶ Improved performance
  - ▶ Reduced training time compared to full training
  - ▶ Leverages pre-trained models
- ZenML:
  - ▶ Automatic caching for faster training runs
  - ▶ User-friendly UI to schedule runs and compare results
  - ▶ Automates pipeline execution based on data dependencies
- GWDG ZenML template:
  - ▶ Runs on KISSKI resources
  - ▶ 35 nodes with 4 A100 80 GB GPUs available
  - ▶ Integrates technologies to reduce resource usage

# Outline

- 1 Willkommen
- 2 ChatAI
- 3 Other Services
- 4 Custom Services
- 5 Community**

# ChatAI Community

- ChatAI ist zwar Open Source, aber wer managed die Weiterentwicklung?
- Unser Ziel ist die offene Community Entwicklung
- Lizenz: GPLv3
- Contributor Agreement (CLA)
- Entwicklung einer Community Roadmap
- Governance
  - ▶ Gemeinsam über Features und Prioritäten entscheiden
  - ▶ Voting Rechte: steering board (of contributors)
- Monatliche Treffen der Community
- Initiale Mailingliste auf GöAID: <https://gwdg.de/hpc/events/goeaid/>

## Verein - KI4D e.V.

*“Der Verein zielt darauf ab, die **Integration und Bereitstellung** von Künstlicher Intelligenz (KI) sowohl in die deutsche Industrie, Forschung als auch für jeden einzelne Haushalt in Deutschland zu fördern und zu erleichtern. Wir streben danach, die Verständnis- und Anwendungsbereiche der KI zu erweitern, um effizientere Geschäftsprozesse, verbesserte Dienstleistungen und einen erhöhten Lebensstandard **für alle** Bürger zu gewährleisten. Unser Ziel ist es, die Barriere für den Zugang zu KI-Technologien zu senken und gleichzeitig sicherzustellen, dass diese Technologien verantwortungsbewusst und ethisch eingesetzt werden.”*

# Verein

- Arbeitstitel: KI für Deutschland e.V. - KI4D-ev
- Mitglieder können Institutionen, Industrie oder andere Vereine sein
- Zusammenarbeit mit den KI Servicezentren
  - ▶ Gemeinsame Software Lösungen
  - ▶ Aufwände zusammenführen
- Open Source mit Community Governance
  - ▶ Öffentlich für neue Feature abstimmen
- Meldet euch bei Interesse zur Gründung gerne hier an:  
<https://listserv.gwdg.de/mailman/listinfo/ki4d-ev>



# Summary

- Free, private LLM service with API  
`chat-ai.academiccloud.de`
- Many more AI services available
- Community meeting GöAID  
`gwdg.de/hpc/events/goeaid`
- More technical details via our paper  
`arxiv.org/abs/2407.00110`



For all available KISSKI services go to:  
`kisski.gwdg.de/leistungen/services/`



GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung