GWDG / KISSKI
Julian.Kunkel@gwdg.de

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 wichig
- 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

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`

Motivation

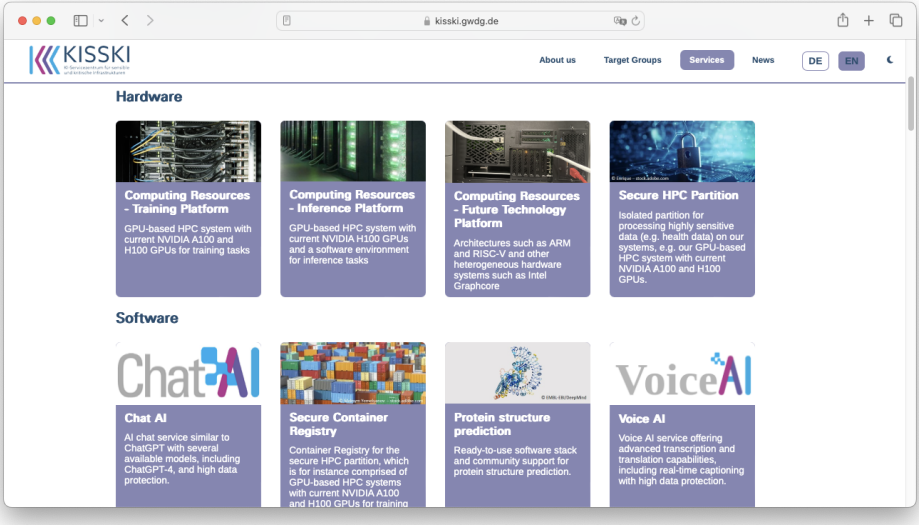
We provide an LLM-Service with

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

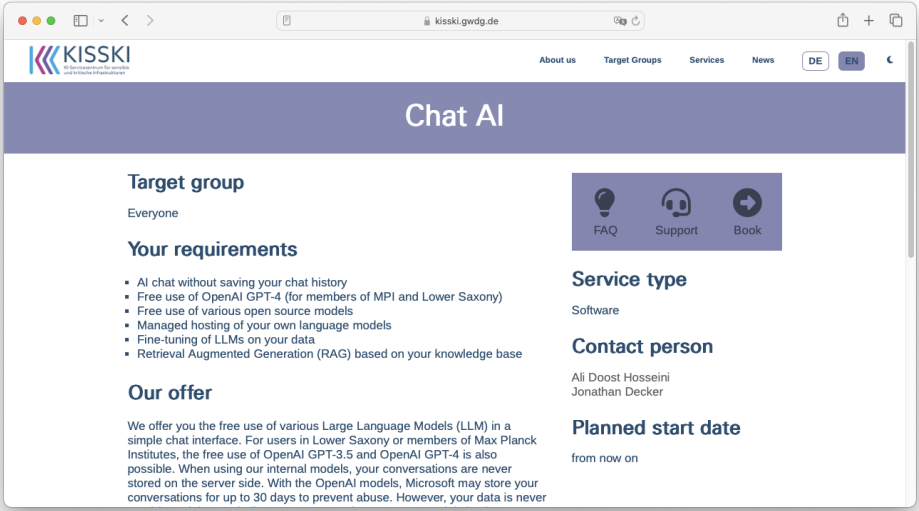
Result: ChatAI

https://chat-ai.academiccloud.de

Access to the AI-Service-Center



Access to the AI-Service-Center



ChatAI

KISSKI GWDG

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:

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/`
  - 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

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

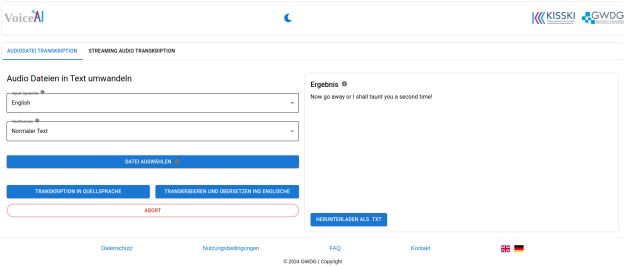
- A code completion service provided by GWDG through the KISSKI project.
- Assist with editing, generating, fixing and commenting code.
- Operates in Visual Studio Code via the Continue plugin.
- Accesses the same LLMs that ChatAI has access to.
- Same security and data privacy that ChatAI provides.

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.

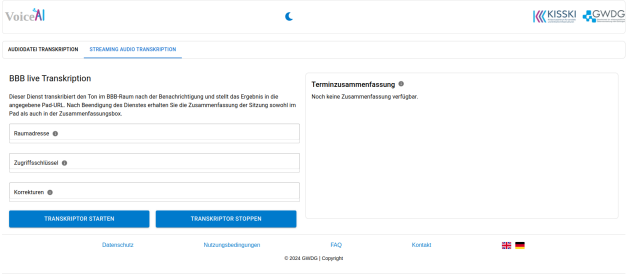
VoiceAI: Transcription and Translation of Uploaded Audio

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



VoiceAI: Live BigBlueButton (BBB) Captioning

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



Protein AI

- **Code:** Colabfold
  - **Rich databases**
  - **Fast and sensetive MMseqs2 method**
  - **Comparative results to Alphafold2** (Full system)

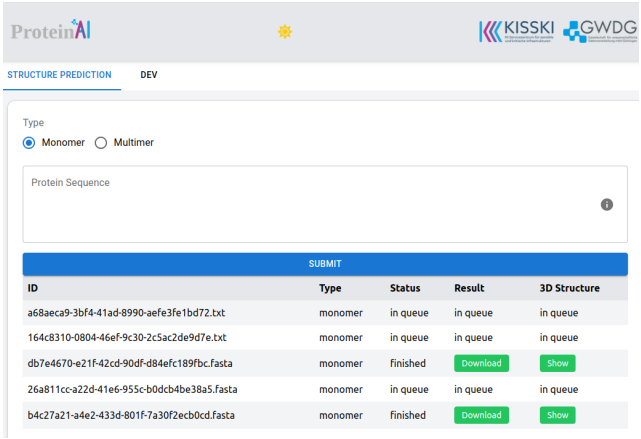


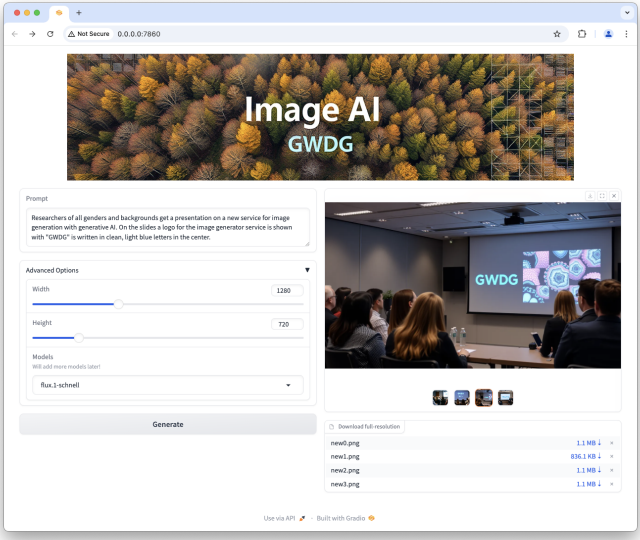
Image AI

- Text zu Bild Generierung
- erstes Modell `FLUX.1-schnell`[a][b]
- OpenAI-kompatibler API-Server auf der KISSKI-Inferenzplattform
- (optionale Funktion) Bild-zu-Bild

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

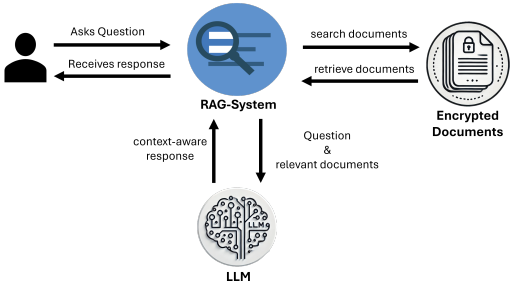
[b] prompt: "A high performance computing cluster. In the Background a cat sitting on top of the HPC cluster and holding a sign that says 'Coming soon!'. At the top 'GWDG', in clean, simple, light blue letters. In the center of the image 'Image AI' in clean, simple, light blue letters."

Image AI


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
GWDG
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

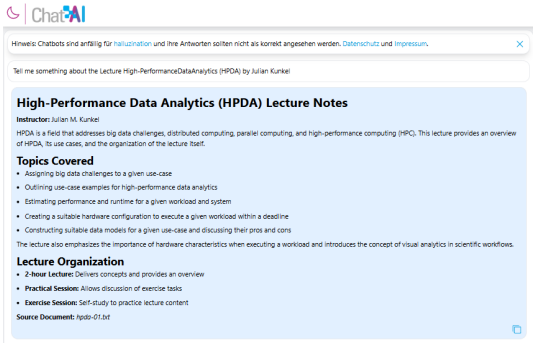
RAG Integration: Overview

- 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.



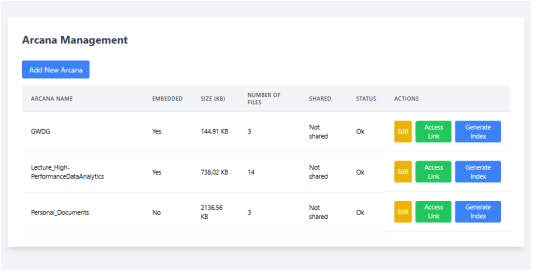
RAG Integration: Middleware and Chat Interface

- 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.



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.



Outline

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

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/`

KISSKI
KI-Servicezentrum für sensible und kritische Infrastrukturen

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung