GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
IN PUBLICA COMMODA SEIT 1737

Lorenz Glißmann

KI-Dienste an der Georg-August-Universität Göttingen

Ablauf

- Einstieg und Überblick
- Mehrere Workshop-Stationen zum Kennenlernen der Dienste
  - Station kennenlernen
  - Austausch über Erfahrungen

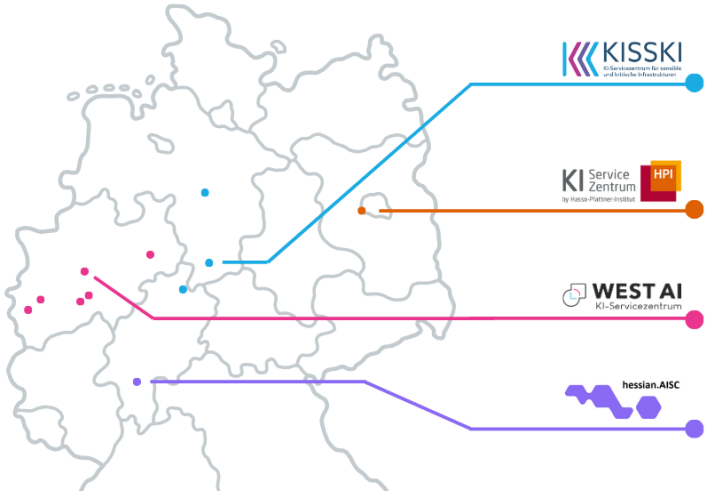
Für diesen Workshop gilt ein *Workshop-Du*!

KISSKI als KI-Servicezentrum

Stärkung der Marke **„AI made in Germany“**

- Forschung und Aufbau von **KI-Expertise in Deutschland**
  - Skalierbarkeit von KI-Methoden
  - KI für große Datenmengen/KI-Modelle
  - KI auf unterschiedlicher/neuartiger Hardware
- Transfer in die Praxis
  - Unentgeltliche Durchführung von **Pilotprojekten**
  - Vernetzung von Wissenschaft und Wirtschaft zum beiderseitigem Vorteil
  - Befähigung von KMUs & Start-Ups zu KI-bezogenen Innovationen

Die 4 national BMFTR-geförderten KI-Servicezentren



KISSKI
KI-Servicezentrum für sensible und kritische Infrastrukturen

Uni Göttingen | GWDG | CIDAS | UMG | aQua | LUH | L3S | LiFE | LUIS | MHH | Fraunhofer IEE
Sensible und kritische Infrastrukturen
Medizin & Energie

KI Service Zentrum by Hasso-Plattner-Institut
HPI

Hasso-Plattner-Institut
Bildungs- und Beratungsangebote,
Einsatz von KI in Wirtschaft &
Gesellschaft

WEST AI
KI-Servicezentrum

Uni Bonn | Fraunhofer IZB | Forschungszentrum Jülich | RWTH Aachen | TU Dortmund | Uni Paderborn
KI-Hardware, KI-Beratung, KI-Schulungen sowie multimodale und transferierbare KI-Modelle

hessian.AISC

TU Darmstadt
Erklärbarkeit, Generalisierbarkeit
und kontextuelle Anpassung

Generative KI-Dienste für Endanwender und gehostete Dienste

- KI Lösungen nutzbar im Browser
  - Chat AI, Coco AI, Protein AI, Voice AI, Image AI, RAG
  - API Zugang zu KI Modellen
- Nutzungsverträge für Open AI Nutzung
- Preismodelle für OpenAI Modelle

chat-ai.academiccloud.de/

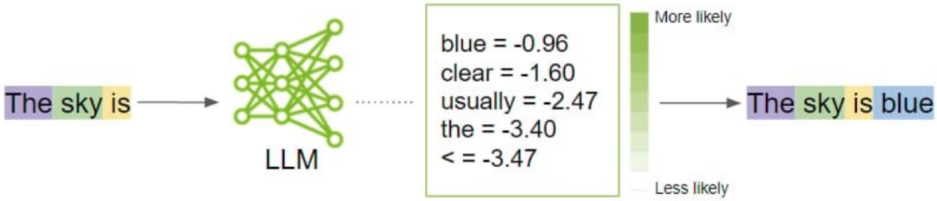
voice-ai.academiccloud.de/

protein-ai.academiccloud.de/

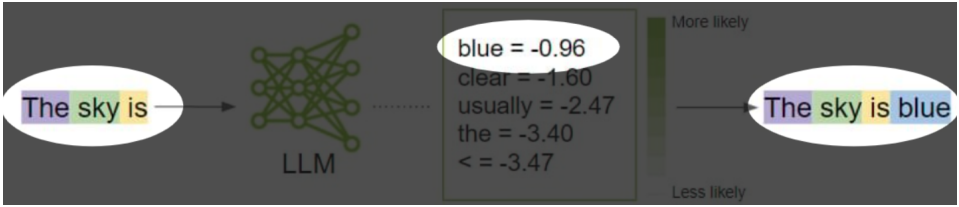
docs.hpc.gwdg.de/services/coco/index.html

Wie arbeiten LLMs?

- Im LLM wird mit Zahlen als Repräsentationen gearbeitet
- Zahlen werden wieder in Tokens/Wörter umgewandelt



Wie arbeiten LLMs?



Chat AI

- Derzeit 18 Open Source LLMs (+ 10 kostenpflichtige Modelle)
- **Konfigurierbar**
- **Import/Export** Konversationen
- Einfaches Sharing per Link
- Anzeige der Modelverfügbarkeit
- https://chat-ai.academiccloud.de

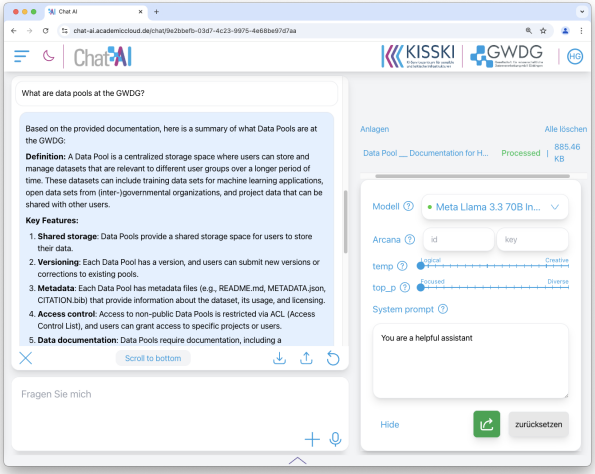


Image AI

- Text-to-image Generierung
- Erstes Model `FLUX.1-schnell`
- Basiert auf OpenAI-kompatiblem API Server
- Optional: Image-to-image
- Mini-Livedemo!

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."

Retrieval-Augmented-Generation (RAG)

- = LLM erhält zusätzliche Daten
- Wissensbasis: "Vektordatenbank"
  - In Abschnitte zerteilte Dokumente
  - Ähnlichkeitsbasierte Suche (gegen Nutzereingabe)
  - Besonders gut funktionieren FAQ-artige Daten
- Ergebnis: ChatBot der auf Wissen zurückgreifen kann

Retrieval-Augmented-Generation (RAG)

- = LLM erhält zusätzliche Daten
- Wissensbasis: "Vektordatenbank"
  - In Abschnitte zerteilte Dokumente
  - Ähnlichkeitsbasierte Suche (gegen Nutzereingabe)
  - Besonders gut funktionieren FAQ-artige Daten
- Ergebnis: ChatBot der auf Wissen zurückgreifen kann

Was braucht man für einen RAG-basierten Chatbot?

1. Dokumente
2. Aufteilung (Chunking)
3. RAG-System $\rightarrow$ ChatAI

In Chat AI heißen RAG-Systeme *Arcanas*

AI Cards

- Problem: KI-Nutzung (in der Wissenschaft) transparent machen
- Lösung: Strukturierte Templates "AI-Cards"
  - Nutzer kann Bereiche/Umfang angeben
  - Funktioniert mit dem Schriftsatzsystem LaTeX
- AI-Cards: https://ai-cards.org/, https://beta.ai-cards.org/
- Entwickelt an der Uni Göttingen, https://gipplab.org/

Wahle, Jan Philip, et al. "Ai usage cards: Responsibly reporting ai-generated content." 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 2023.

Stationen

1. Chat AI
2. Image AI
3. RAG-basierte Chatbots entwickeln
4. AI Cards

Zusammenarbeiten (2-3 Personen) ist explizit erwünscht!