

Offener Zugang zu Privater KI

LLM-Service mit HPC

Christian Boehme, Jonathan Decker, Julian Kunkel



GWDG, Uni Göttingen und KISSKI

■ GWDG

- ▶ IT-Service Anbieter für Uni Gö und MPG
- ▶ Nationales Hochleistungsrechnen, Rang 156 weltweit

■ Georg-August-Universität Göttingen

- ▶ Zweitgrößte Uni in Niedersachsen

■ KISSKI

- ▶ 1 der 4 KI Servicezenter in Deutschland
- ▶ BMBF-gefördert seit 10/2022
- ▶ Verschiedene Dienste via <https://kisski.gwdg.de/>



KI-Servicezentren

WestAI

Dortmund/Bonn/Jülich/Aachen/Paderborn

hessian AI Service Center

Darmstadt

KISSKI

Hannover/Göttingen/Kassel

KI-Servicezentrum Berlin Brandenburg

Hasso-Plattner-Institut

Zugang zu KI-Service-Zentren

KISSKI
KI-Servicezentrum für sensible und kritische Infrastrukturen

Über uns Zielgruppen Leistungen Aktuelles DE EN

KISSKI
KI-Servicezentrum für sensible und kritische Infrastrukturen

+++ AKTUELL +++
How to KISSKI
Lernen Sie, die KISSKI Infrastruktur zu nutzen. Am 08.07.2024 bieten wir im Rechenzentrum Göttingen zwei Kurse an, die Ihnen einen Einstieg in die Nutzung der HPC-Ressourcen des Servicezentrums ermöglichen.
[mehr]

+++ AKTUELL +++
Konferenz der deutschen KI-Servicezentren (KonKIS 24)
Am 18. und 19.09.2024 findet die erste Konferenz der deutschen KI-Servicezentren

KISSKI
KI-Servicezentrum für sensible und kritische Infrastrukturen

Über uns Zielgruppen Leistungen Aktuelles DE

Zielgruppen

Alle

Ihre Anforderungen

- KI-Chat ohne Speicherung Ihres Chatverlaufs
- Kostenlose Nutzung von OpenAI GPT-4 (nur für Niedersachsen und MPI)
- Kostenlose Nutzung von diversen Open-Source-Modellen
- Managed hosting Ihrer Sprachmodelle
- Finetuning von LLMs auf ihren Daten
- Retrieval-Augmented Generation (RAG) auf ihren Dokumenten

Unser Angebot

Wir bieten Ihnen die kostenlose Nutzung diverser Large Language Models (LLM) in einer einfachen Chat-Oberfläche an. Für Nutzer*Innen in Niedersachsen oder Mitglieder von Max-Planck Instituten ist auch die kostenlose Nutzung von OpenAI GPT-3.5 und OpenAI GPT-4 möglich. Bei der Nutzung unserer internen Modelle werden nie serverseitig Ihre Konversationen gespeichert. Bei den OpenAI Modellen kann Microsoft Ihre Unterhaltungen bis zu 30 Tage lang speichern, um Missbrauch zu verhindern, jedoch nicht zu Trainingszwecken oder Ähnliches. Auf Anfrage ist auch das Hosting Ihrer eigenen Modelle möglich.

Außerdem bieten wir Finetuning Ihrer Modelle und Retrieval-Augmented Generation (RAG), eine effektive und

Art des Services
Software

Ansprechpartner:in
Ali Doost Hosseini
Jonathan Decker

geplanter Starttermin
2024

Chat AI

Table of contents

- 1 Chat AI Übersicht
- 2 Nutzungsstatistiken
- 3 Weiteres

Problemstellung

- Forschende, Studierende möchten LLMs nutzen
- Starke LLMs benötigen viele teure Grafikkarten
- OpenAI, Google bieten LLM Service an, keine Datenschutzgarantie

Problemstellung

- Forschende, Studierende möchten LLMs nutzen
- Starke LLMs benötigen viele teure Grafikkarten
- OpenAI, Google bieten LLM Service an, keine Datenschutzgarantie

Bereitstellung eines LLM-Service mit

- Skalierbarer Infrastruktur
- Open Source Basis
- Strengem Datenschutz

Problemstellung

- Forschende, Studierende möchten LLMs nutzen
- Starke LLMs benötigen viele teure Grafikkarten
- OpenAI, Google bieten LLM Service an, keine Datenschutzgarantie

Bereitstellung eines LLM-Service mit

- Skalierbarer Infrastruktur
- Open Source Basis
- Strenge Datenschutz



Resultat: Chat AI

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



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 ⓘ

Komponenten

- Chat Web Interface
 - ▶ Eigenentwicklung basierend auf React und Vue.js
- Middleware
 - ▶ Eigenentwicklung basierend auf Kong
- HPC Scheduler
 - ▶ Eigenentwicklung basierend auf Slurm
- HPC Compute Ressourcen
 - ▶ 12 GPU Server finanziert durch KISSKI
- Sprachmodelle
 - ▶ Top Open Weights Modelle wie Llama3, Mixtral und Qwen2

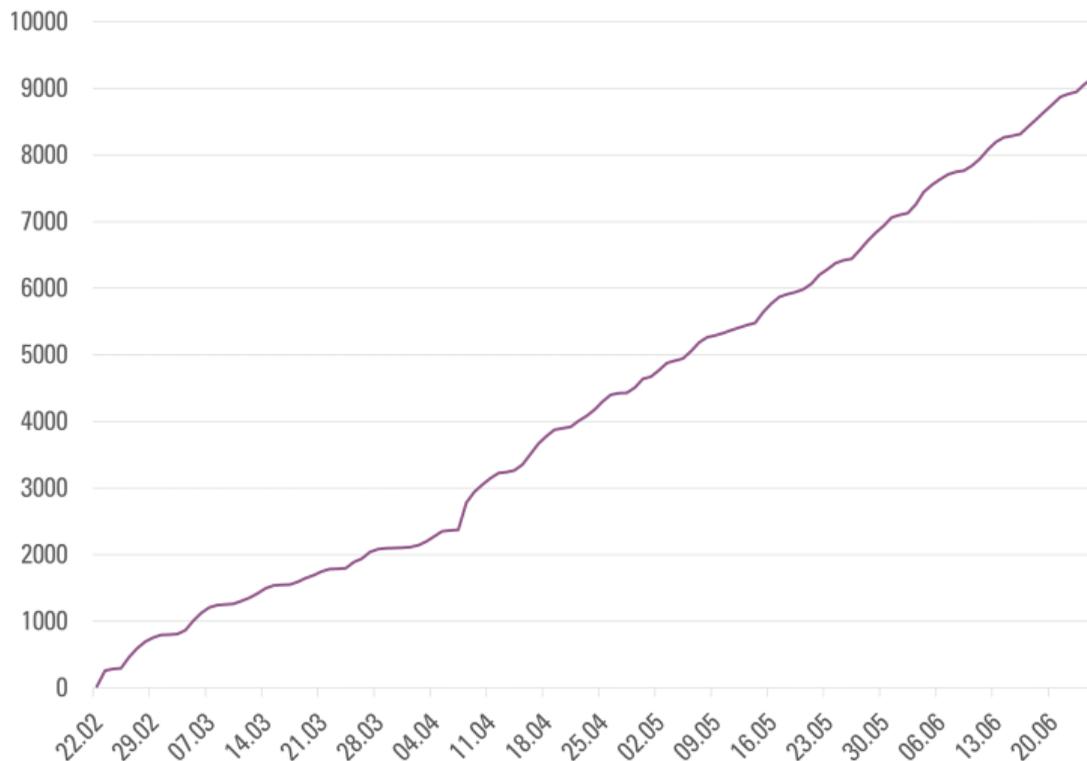
Features

- Strenger Datenschutz
 - ▶ Nutzer Anfragen und Antworten werden nie auf Server gespeichert
- Mächtiges Web Chat Interface
 - ▶ Spracheingabe
 - ▶ Up- und Download von Konversationen
 - ▶ Wahl des LLMs via Dropdown
 - ▶ Systemprompt frei konfigurierbar
 - ▶ Integrierter Zugriff auf ChatGPT4
- API Zugriff via Standard API (OpenAI kompatibel)
- Auswahl an Open Source Modellen

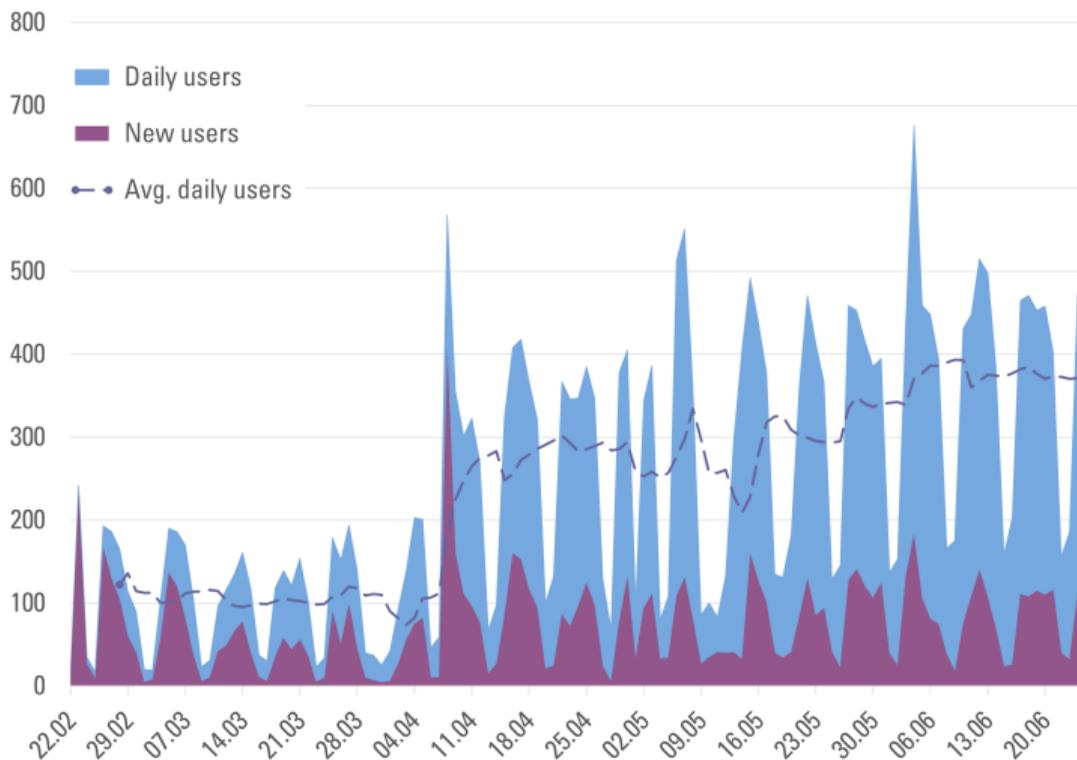
Beitrag zu offenem Ökosystem

- Freier Zugriff auf offene Modelle für alle
 - ▶ Benötigt kostenlosen Academiccloud Account
- API Zugriff möglich
 - ▶ Kostenlos via <https://kisski.gwdg.de/> buchen
 - ▶ Nutzbar in externen Applikationen, e.g., SillyTavern, HAWKI
- ChatGPT4 Zugriff kostenlos für Nutzer in Niedersachsen und aus MPG
 - ▶ Vertrag für ChatGPT über uns möglich

Gesamt Verschiedene Nutzer

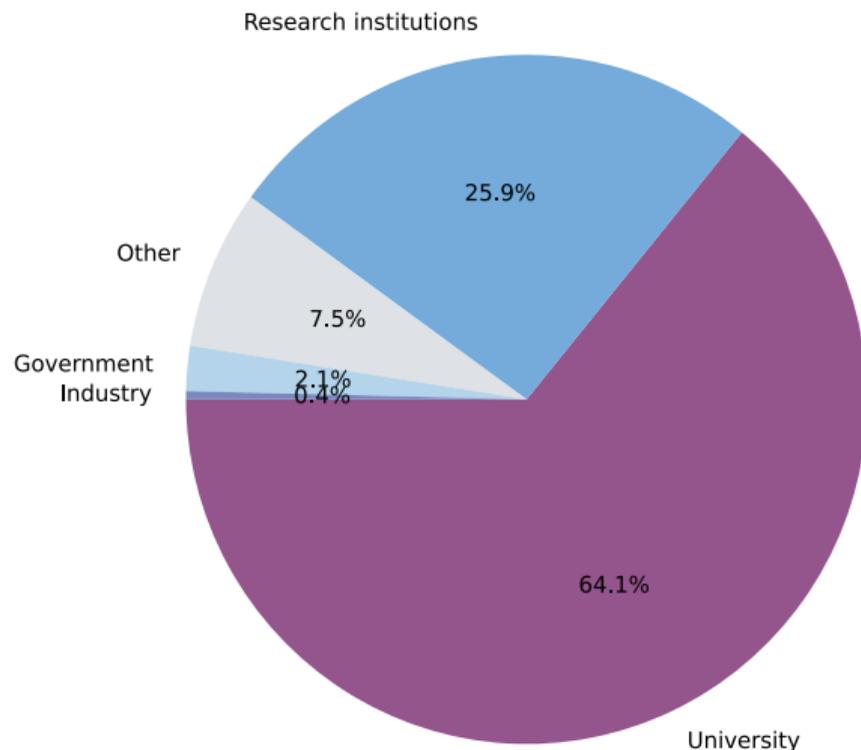


Tägliche wiederkehrende und neue Nutzer

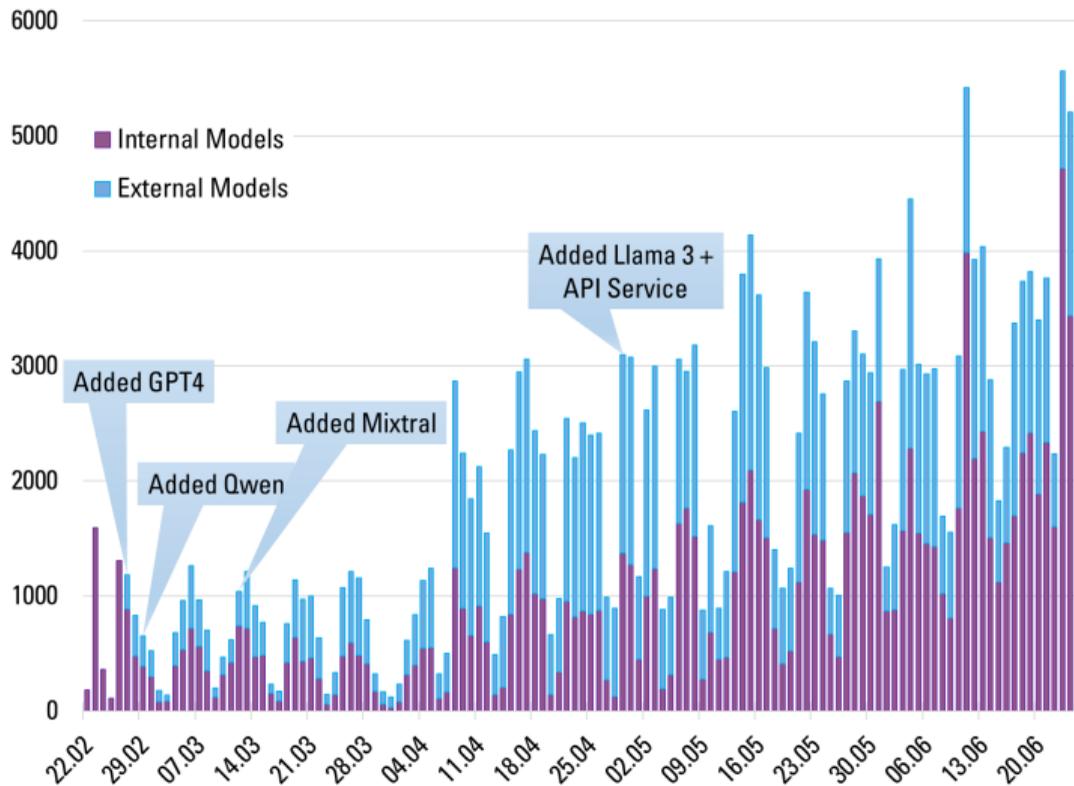


Nutzer nach Demographien

- Nutzer nach Demographie ohne Volumen
- 30% ist Uni Göttingen
- Insgesamt über 160 verschiedene Unis
- Viele MPG Institute unter *Research Institutions*
- Bei Industrie Nutzer meist nur ein API Key pro Firma



Tägliche Nutzer Anfragen

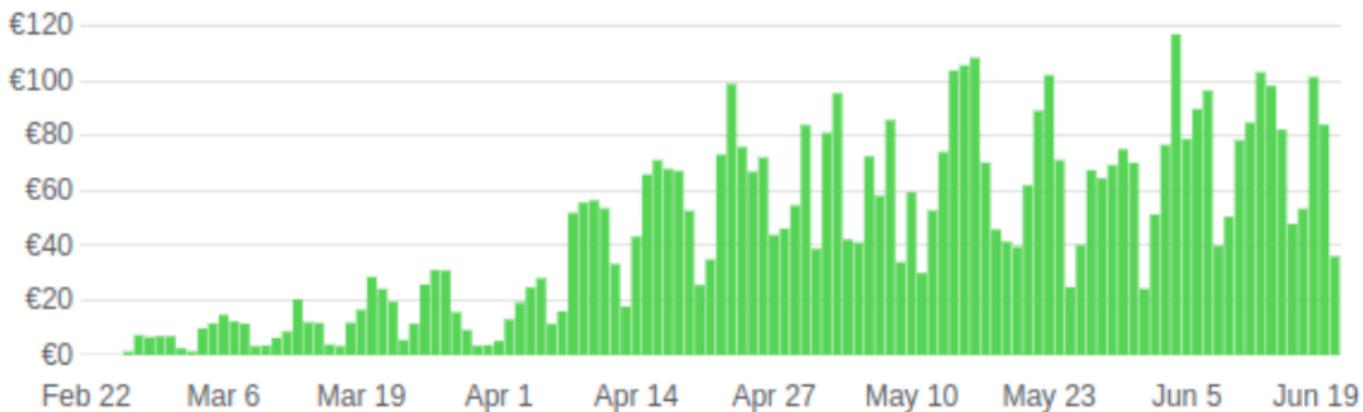


Ressourcenbedarf

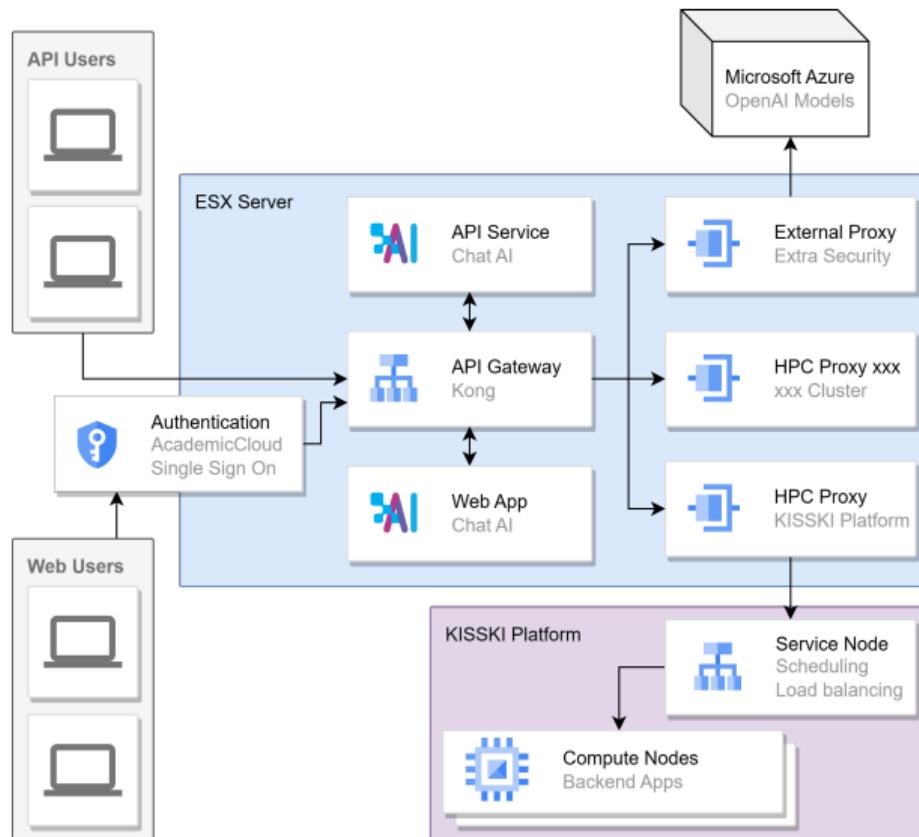
- 70B LLMs benötigen je 4 NVIDIA H100 Grafikkarten
 - ▶ Mit reduzierter Präzision (Quantisierung) 2
 - ▶ Stabilität der quantisierten Modelle reicht aktuell nicht aus
- Aktive Instanzen via Anfragen Volumen automatisch skaliert
 - ▶ Start eines neuen Modells benötigt bis zu 10 Minuten
 - ▶ Daher 1 permanente Instanz pro Modell erforderlich für akzeptable Antwortzeit
- 12 Server je 4 H100 Gesamt verfügbar
- Web Interface läuft in on-premise Cloud
- An Entwicklung, Betrieb, Support sind 5 Personen beteiligt

Kosten für OpenAI Zugriff

- Zugriff zu ChatGPT4 ist kostenlos für
 - ▶ Nutzer in Niedersachsen
 - ▶ MPG Mitglieder
- Bisher über 5000€ an Azure bezahlt
 - ▶ Davon 2000€ im Mai



Architektur



Scalable AI Accelerator - SAIA Plattform und Ökosystem

- Code completion in VS Code via API
- Spracherkennung
 - ▶ Whisper Modell
 - ▶ Big-Blue-Button Integration in Arbeit
- Fine-Tuning
 - ▶ 35 Server mit je 4 A100 80GB GPUs via KISSKI verfügbar
- Beratung zum Einsatz von KI
 - ▶ Eigene Daten in LLM via RAG
 - ▶ Datenschutz-konforme Strategien
- Chat AI Erweiterungen in Arbeit
 - ▶ Support für Vision Language Models (VLM)
 - ▶ Integration von Bild Generation, e.g., Stable Diffusion
- Nach ISO 27001 zertifiziert

Abschluss

- Freier, privater LLM Chat Service mit API
- Viele weitere Services zur Auswahl
- <https://chat-ai.academiccloud.de>

Chat AI

KISSKI
KI-Servicezentrum für sensible
und kritische Infrastrukturen

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Verteilung alle Nutzer aus Universitäten

