

Free Access to Private AI

AI-Services on the HPC

H. Kirchner, A. Doosthosseini, J. Decker, J. Hördt, N. Lux, C. Mandal, L. van Niekerk, M. H. Biniaz, S. Keshtkar, J. Kunkel



GWDG and KISSKI

■ GWDG

- ▶ IT-service provider for Uni Göttingen and MPG
- ▶ System “Emmy+”:
 - 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



KI-Servicezentren

WestAI

Dortmund/Bonn/Jülich/Aachen/Paderborn

hessian AI Service Center
Darmstadt

KISSKI

Hannover/Göttingen/Kassel

KI-Servicezentrum Berlin Brandenburg
Hasso-Plattner-Institut

Table of contents

- 1 ChatAI Overview
- 2 Technical Details
- 3 Beyond ChatAI
- 4 Custom Services
- 5 Community

Problem Statement

- Researchers, students want to use LLMs
- Powerful LLMs require many expensive GPUs
- OpenAI, Google offer LLM services, no data privacy guarantees

Problem Statement

- Researchers, students want to use LLMs
- Powerful LLMs require many expensive GPUs
- OpenAI, Google offer LLM services, no data privacy guarantees

Provisioning of an LLM-Service with

- Scalable infrastructure
- Open source base
- Built-in data privacy

Problem Statement

- Researchers, students want to use LLMs
- Powerful LLMs require many expensive GPUs
- OpenAI, Google offer LLM services, no data privacy guarantees

Provisioning of an LLM-Service with

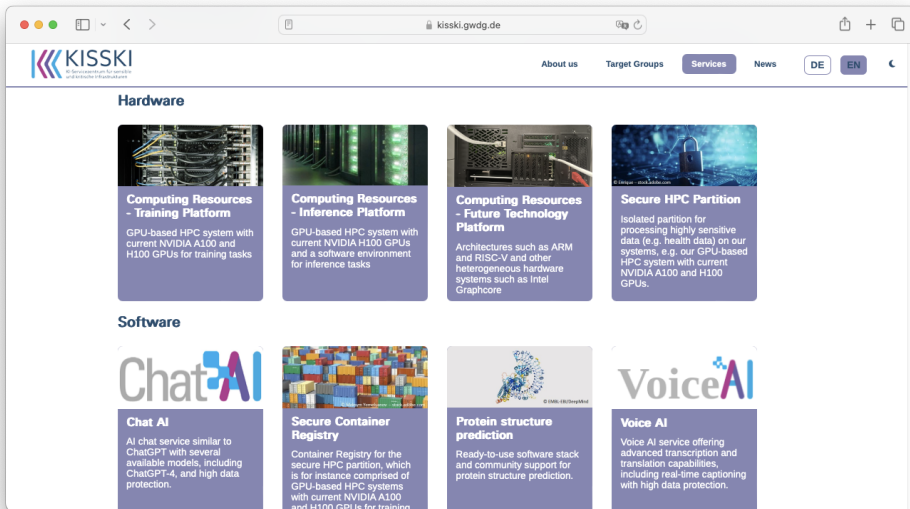
- Scalable infrastructure
- Open source base
- Built-in data privacy



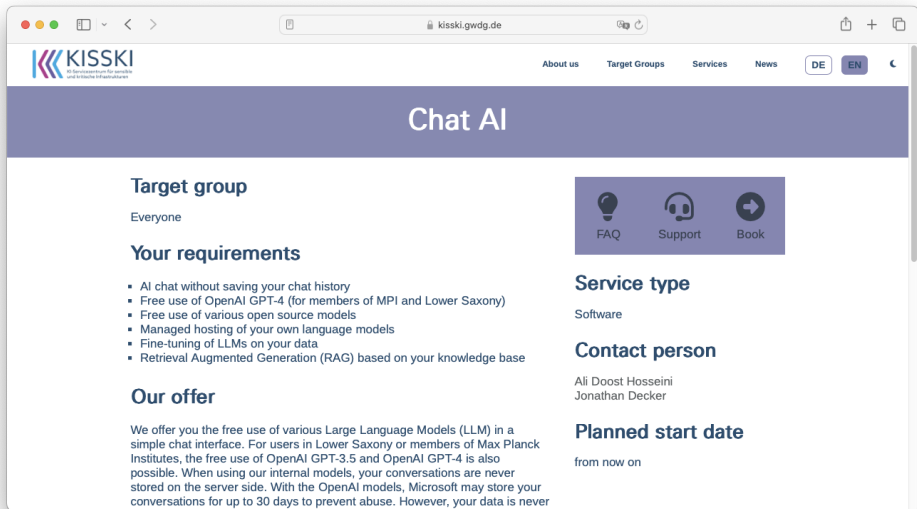
Result: ChatAI

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

Access to the AI-Service-Center



Access to the AI-Service-Center



The screenshot shows a web browser window displaying the KISSKI website. The browser's address bar shows 'kisski.gwdg.de'. The website has a navigation bar with links for 'About us', 'Target Groups', 'Services', 'News', and language selectors 'DE' and 'EN'. The main heading is 'Chat AI'. Below this, there are four sections: 'Target group' (Everyone), 'Your requirements' (a list of conditions for free use), 'Our offer' (a paragraph about the service), and 'Service type' (Software). To the right of these sections are three icons: a lightbulb for 'FAQ', a headset for 'Support', and a right arrow for 'Book'. Below the 'Service type' section, there is a 'Contact person' section listing 'Ali Doost Hosseini' and 'Jonathan Decker', and a 'Planned start date' section stating 'from now on'.

KISSKI
AI 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 PyTorch neural network class from the `nn.Module` class.



Ask me



Model ⓘ

Meta LLaMA 3 70B Instruct



Advanced options ⓘ

Components

- Chat web interface
 - ▶ Own development based on React and Vitejs
- Middleware
 - ▶ Own development based on Kong
- HPC scheduler
 - ▶ Own development based on Slurm
- HPC compute resources
 - ▶ 12 GPU Server financed via KISSKI
- LLMs
 - ▶ Top open weights models such as Llama3.1, Mixtral and Qwen2

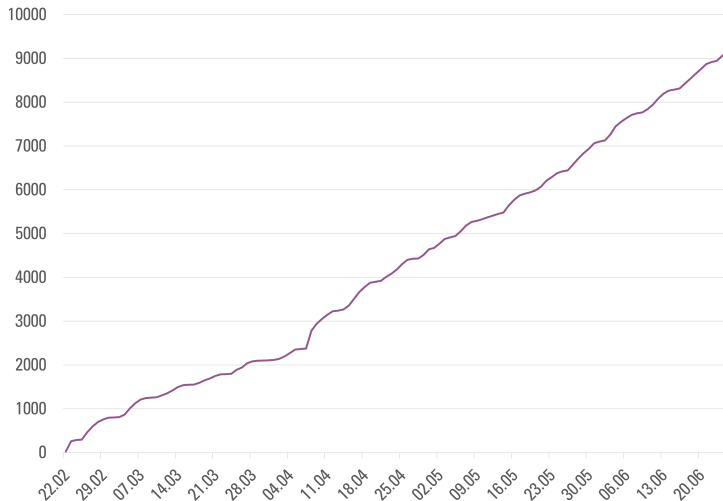
Features

- Built-in data privacy
 - ▶ 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

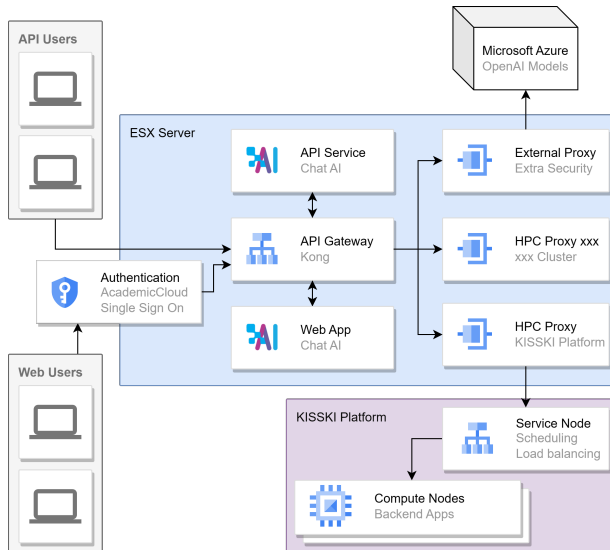
Contribution to an Open Ecosystem

- 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

Total Distinct Users

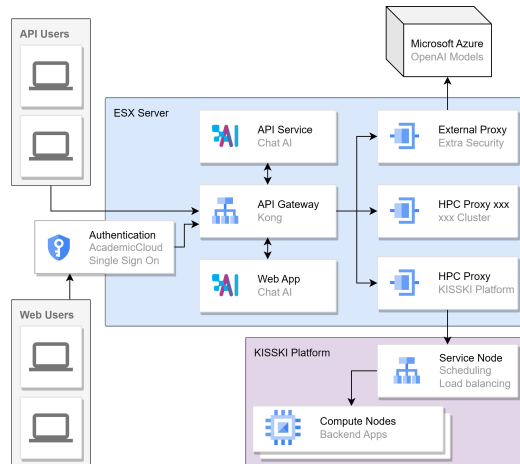


Architecture



Architecture - Internal Workings

- Proxy handles authentication
- Gateway handles all internal routing
- HPC Proxy maintains SSH tunnel with HPC
- Uses SSH ForceCommand to act as circuit breaker
- vLLM runs in Slurm jobs



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 shows the VoiceAI web interface. At the top, there's 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 section is titled "Audio Dateien in Text umwandeln". It contains two dropdown menus: "Zielsprache" (set to "English") and "Textformat" (set to "Normaler Text"). Below these are three buttons: "DATEI AUSWÄHLEN" (blue), "TRANSKRIPTION IN QUELSPRACHE" (blue), and "TRANSKRIBIEREN UND ÜBERSETZEN INS ENGLISCHE" (blue). A red "ABORT" button is also present. To the right, there's a box labeled "Ergebnis" containing the text "Now go away or I shall taunt you a second time!". At the bottom of this box is a blue button "HERUNTERLADEN ALS TXT". The footer contains links for "Datenschutz", "Nutzungsbedingungen", "FAQ", and "Kontakt", along with a copyright notice "© 2024 GWDG | Copyright" and a German flag icon.

VoiceAI: Live BigBlueButton (BBB) Captioning

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

The screenshot shows the VoiceAI web interface for BBB live transcription. At the top, there's a header with the VoiceAI logo on the left and logos for KISSKI and GWDG on the right. Below the header, there are two tabs: 'AUDIODATEI TRANSKRIPTION' and 'STREAMING AUDIO TRANSKRIPTION', with the latter being selected. The main content area is titled 'BBB live Transkription' and contains a 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 Zusammenfassungsbbox.' Below this text are three input fields: 'Raumadresse @', 'Zugriffsschlüssel @', and 'Korrekturen @'. At the bottom of the main area are two blue buttons: 'TRANSKRIPTOR STARTEN' and 'TRANSKRIPTOR STOPPEN'. To the right of the main area is a sidebar titled 'Terminzusammenfassung @' with the text 'Noch keine Zusammenfassung verfügbar.' At the very bottom of the page, there are links for 'Datenschutz', 'Nutzungsbedingungen', 'FAQ', and 'Kontakt', along with a copyright notice '© 2024 GWDG | Copyright' and a small German flag icon.

Image AI

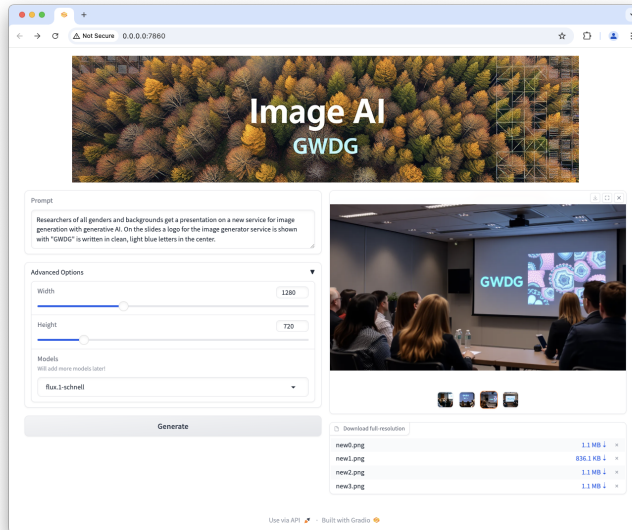
- text-to-image generation
- starting with FLUX.1-schnell^{ab}
- based on OpenAI compatible API server running on KISSKI inference platform
- (optional feature) image-to-image

^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



RAG - PrivateGPT

- Retrieval Augmented Generation (RAG) combines retrieval-based and generative models
 - ▶ Finds the parts of documents most relevant for a query
 - ▶ Quotes those parts into context window
 - ▶ Uses LLM to produce the output
- PrivateGPT: open-source RAG project that preserves user privacy
- Interacts with our ChatAI API for Text Completion and Embedding

The screenshot displays the PrivateGPT web interface. At the top, a purple header bar contains the text "PRIVATE GPT". Below this, the interface is divided into two main sections. On the left, there is a control panel with a "Mode" dropdown set to "RAG", buttons for "Search" and "Basic", and a "Summarize" button. Below these are fields for "Get contextualized answers from selected files.", an "Upload File(s)" button, and a table of "Ingested Files" with one entry: "1786_83762v7.pdf". At the bottom of this panel are buttons for "De-select selected file", "Selected for Query or Deletion", "All files", "Delete selected file", and "Delete ALL files". The right section is the chat area, showing a query "What is the main innovation of the Transformer model?" and a response from the LLM (ollama) stating that the main innovation is its reliance on self-attention. Below the response, two sources are listed: "1. 1706.03762v7.pdf (page 2)" and "2. 1706.03762v7.pdf (page 1)". At the bottom of the chat area are buttons for "Retry", "Undo", "Clear", and a "Submit" button.

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

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

Das erste Treffen im Oktober wird in kürze bekanntgegeben
- 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>

Closing

- Free, private LLM chat 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/

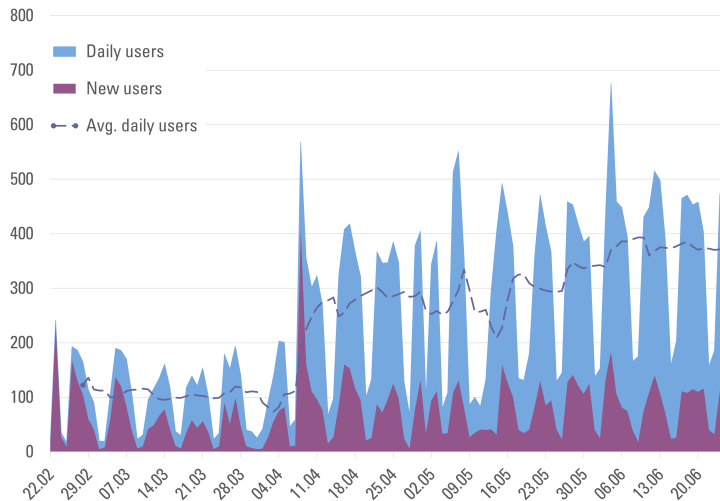


GEFÖRDERT VOM



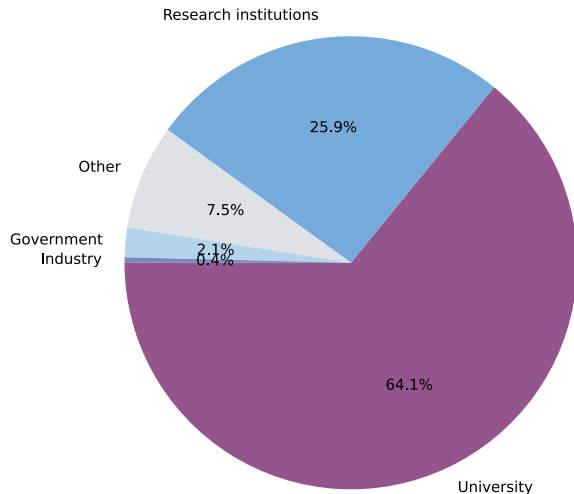
Bundesministerium
für Bildung
und Forschung

Daily Returning Users and New Users

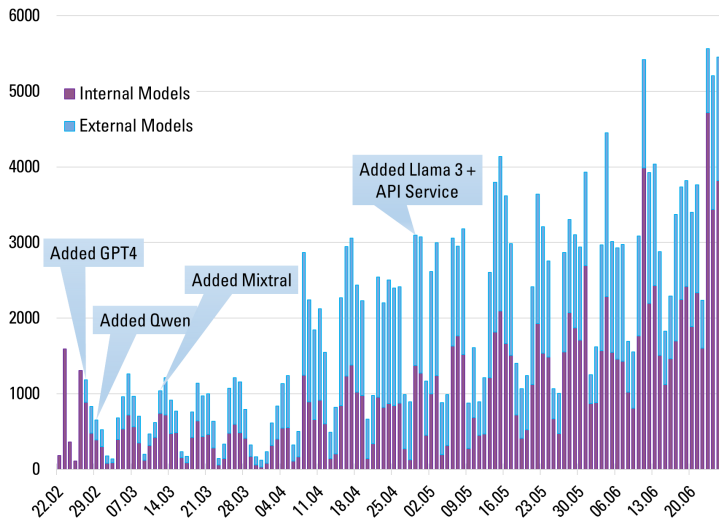


Users by Demographic

- Users by demographic without volume
- 30% is Uni Göttingen
- Total over 160 different Unis
- Many MPG institutes under *Research Institutions*
- Industrial users often only one API key per company



Daily User Requests

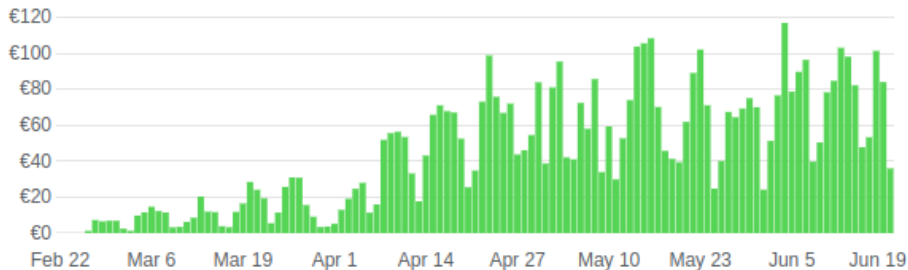


Resource Usage

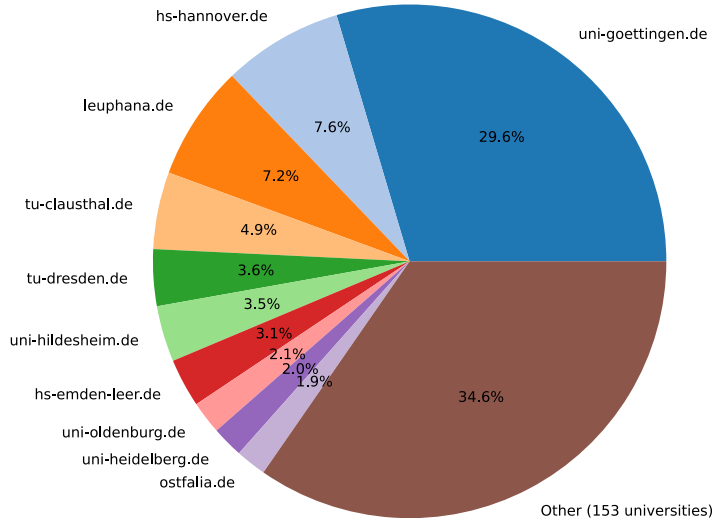
- 70B LLMs require 4 NVIDIA H100 GPUs each
 - ▶ With reduced precision (quantization) 2
- Active instances automatically scale based on request volume
 - ▶ Start of a model takes up to 10 min
 - ▶ One active instance required for acceptable response times
- 12 Server with 4 H100 available in total
- Web interface runs in on-premise Cloud
- 5 employees handle development, operation and support

Costs for OpenAI Access

- Access to ChatGPT4 is free for
 - ▶ User in Lower Saxony
 - ▶ MPG Members
- So far over 9000€ paid to Azure
 - ▶ 2800€ in July
 - ▶ At the moment avg. 6 ct per request



Distribution of All Users by University



Distribution of Users from Universities by Federal State

