

GöAID Kick-Off

Jakob Hördt, Ali Doosthosseini
07.05.2024



Welcome



- Who are we?
 - KI-Servicezentrum für sensible und kritische Infrastrukturen (**KISSKI**)
 - AI service center funded by BMBF
 - Provide AI as a service
 - Offer compute resources on training and inference platforms
 - Consulting
- What is GöAID about?
 - AI developers / MLOps engineers
 - Learning / fun
 - Presentations
 - Open exchange



KISSKI Website: <https://kisski.gwdg.de/>

Potential Topics

- What do **you** do at work?
- AI ecosystems
 - e.g. LangChain, ollama, h2ogpt
 - pytorch, tensorflow tips & tricks
 - Kubernetes & KubeFlow
- AI techniques
 - Model training
 - LLM finetuning
 - Retrieval Augmented Generation (RAG)
- What do **we** do at work?

Chat AI



Chat AI

Write an essay about GWDG

GWDG, also known as the Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen, is a renowned institution in Germany that provides cutting-edge IT services to the scientific community. Established in 1964, GWDG has been at the forefront of technological advancements, empowering researchers and scientists to conduct innovative and groundbreaking research.

One of the primary objectives of GWDG is to provide high-performance computing (HPC) resources to the scientific community. The institution operates several powerful supercomputers, including the HLRN-IV, which is one of the fastest supercomputers in the world. These machines enable researchers to simulate complex phenomena, analyze massive datasets, and model complex systems, thereby driving scientific discovery and innovation. For instance, scientists have used GWDG's HPC resources to study climate change, develop new materials, and understand the behavior of subatomic particles.

In addition to HPC, GWDG offers a range of data management and storage services to support the increasingly data-intensive nature of scientific research. The institution's data center provides secure, reliable, and scalable storage solutions, enabling researchers to store and manage vast amounts of data generated by their research. Furthermore, GWDG's data management services ensure that researchers can easily access, share, and analyze their data, facilitating collaboration and accelerating the discovery process.

GWDG also plays a critical role in promoting digital literacy and providing training and education to the scientific community. The institution offers a range of courses, workshops, and tutorials on topics such as programming, data analysis, and visualization, ensuring that researchers have the necessary skills to effectively utilize advanced computing resources. Moreover, GWDG's experts provide individualized support and consulting services, helping researchers to optimize their workflows and overcome technical challenges.

Another significant aspect of GWDG's work is its commitment to open science and open data. The institution strongly advocates for the principles of transparency, reproducibility, and collaboration, recognizing that open research practices are essential for advancing scientific knowledge. GWDG's open data repository, RADAR, provides a platform for researchers to share their data, making it accessible to the broader scientific community and fostering a culture of collaboration and reuse.

GWDG's impact extends beyond the scientific community, with its work having significant societal and economic implications. The institution's support for research in fields such as climate modeling, medicine, and materials science has the potential to drive innovation and improve people's lives. Moreover, GWDG's expertise in data management and analytics can be applied to a range of industries, from healthcare to finance, driving economic growth and development.

In conclusion, GWDG is a vital institution that plays a critical role in advancing scientific research and innovation in Germany and beyond. Through its provision of high-performance computing resources, data management services, and training and education, GWDG empowers researchers to conduct groundbreaking research and drive discovery. The institution's commitment to open science

Ask me

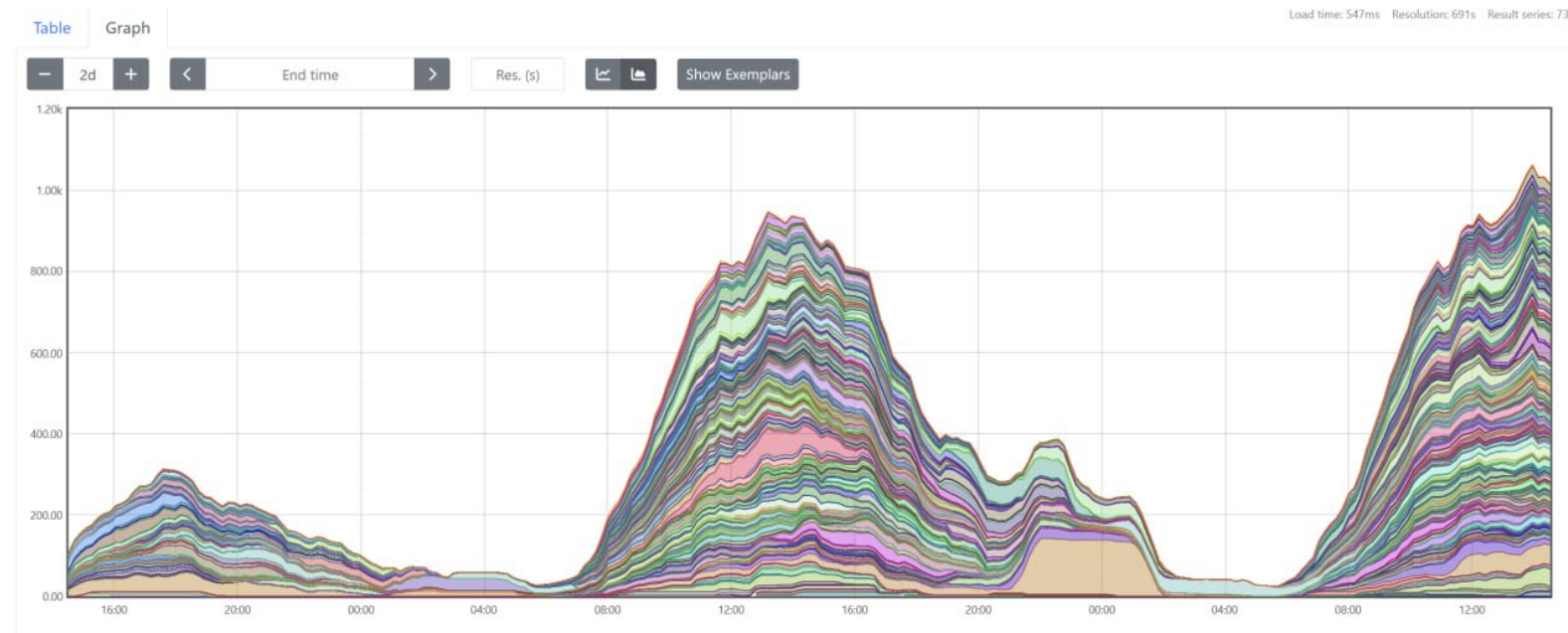
+ 🎤

Model ⓘ Meta LLaMA 3 70B Instruct ▾

Advanced options ⓘ

Chat AI

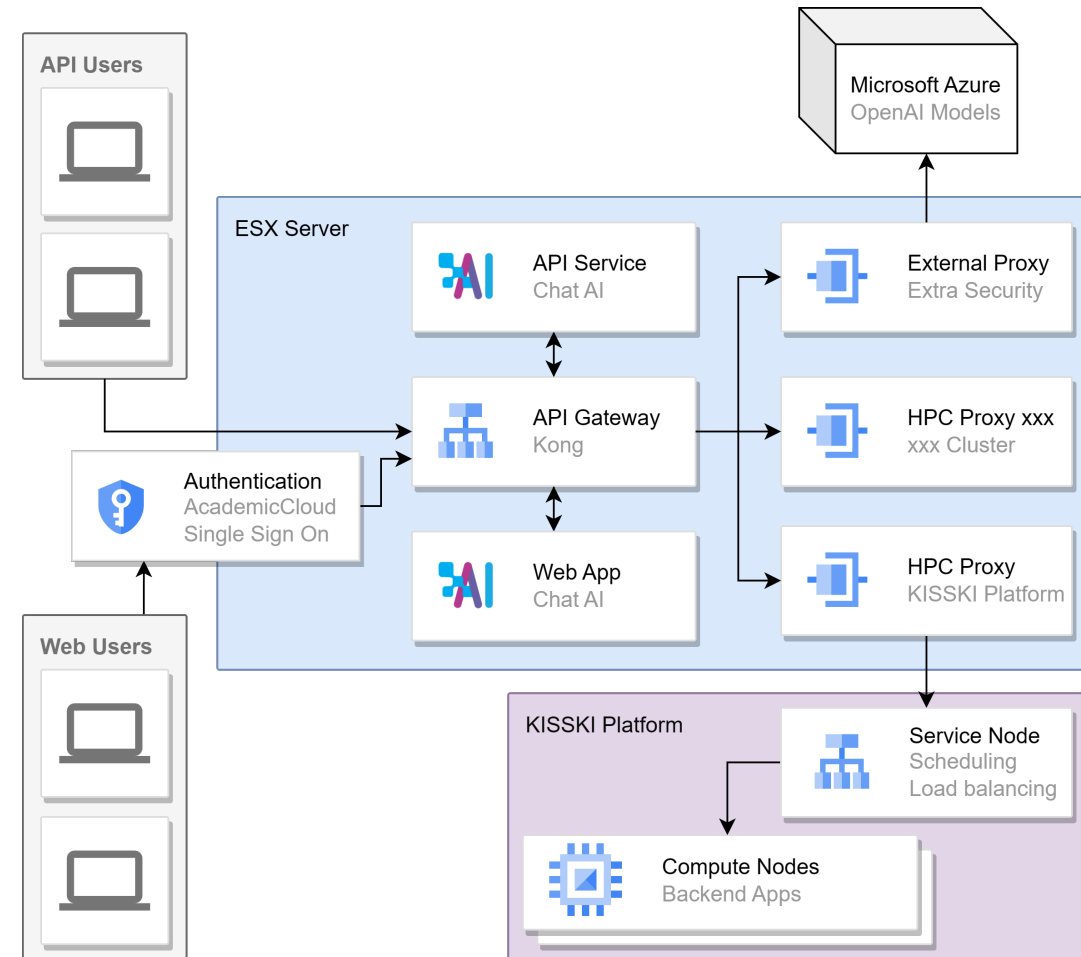
- 3.5K+ individual users
- 3K messages per day
- ~30% internal models
- We offer:
 - Intel Neural Chat 7B
 - Mixtral 8x7B Instruct
 - Qwen 1.5 72B Chat
 - Meta Llama 3 Instruct
 - Your custom models :-)
- Features: custom system prompts, upload text files, save/load chat history



Usage statistics of chat-ai.academiccloud.de

Chat AI – How we did it

- Goals
 - Multiple selectable models
 - Custom models
 - GPU nodes in cluster
 - No data stored on any server
- Methods
 - Use SSH forced-command
 - Use slurm, vLLM in backend
 - OpenAI-compatible API
- Next steps
 - Finetuning platform



SillyTavern Project

- Presented at:
 - Infotage
 - Science goes city
 - IdeenExpo
- Lots of interest from families and younger people
- Experience was useful for Chat AI

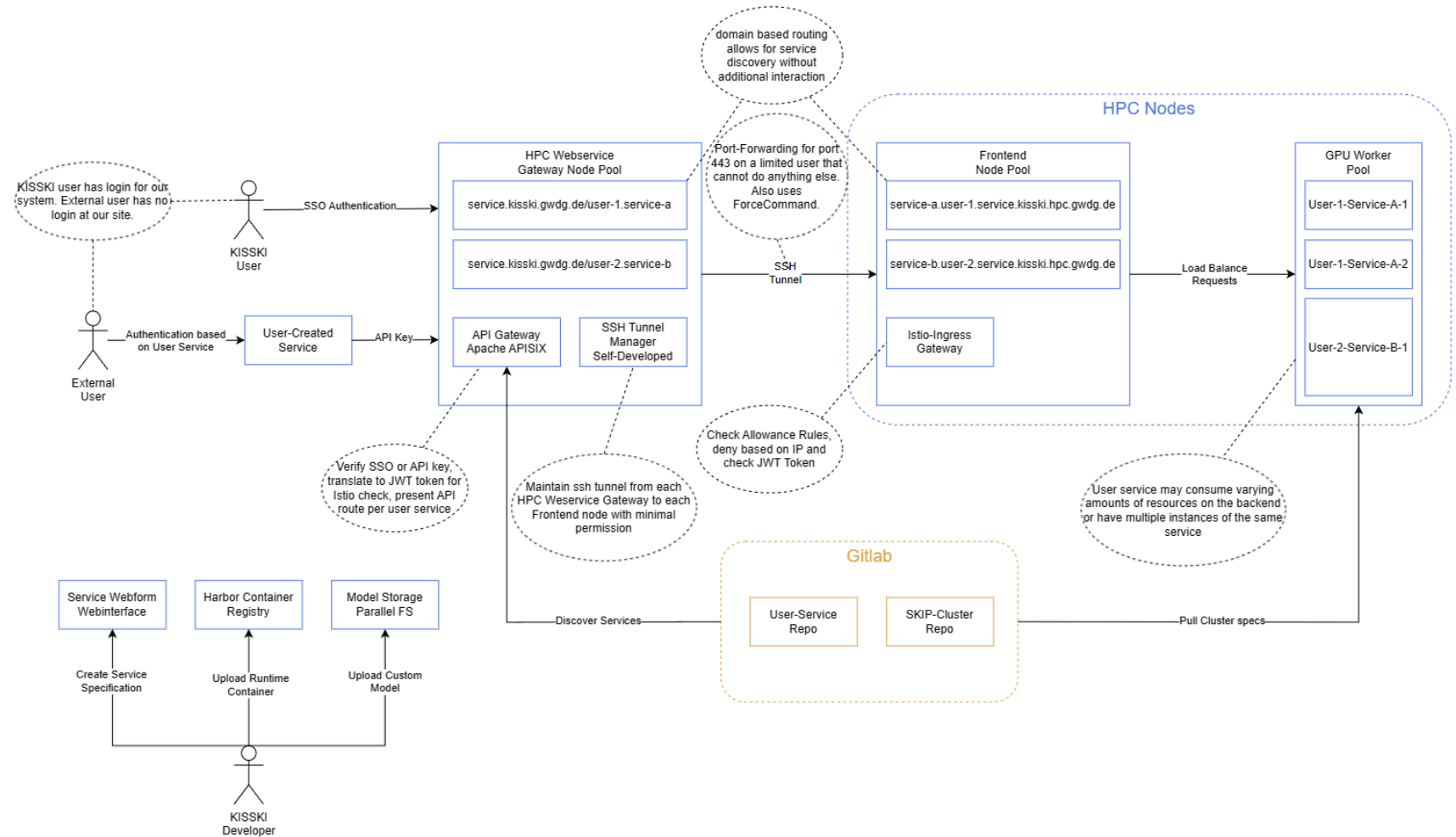


Source: InfoTage Uni Göttingen

SillyTavern Project: <https://sillytavern.app/>

Scalable Kubernetes Inference Platform (SKIP)

- Similar structure to Chat AI, but with Kubernetes clusters
- Better redundancy, scalability



Whisper Service



- Whisper large for automatic language detection and translation
- Automatic video subtitling (maybe as separate service)
- Prototypes are ready, work needed to put into production

Open Discussion



- Let's talk about you...
- What topics are you more interested in?
- Would you like to give a talk about your experiences?
 - Contact Jakob: jakob.hoerdt@uni-goettingen.de
- General Feedback