

Julian Kunkel

Secure Applications for Health Providers and Research-Education

Agenda

1 KISSKI Offer

2 Use Case: Doctor’s letter

3 Use Case: AI in Radiology

4 Secure HPC

5 TRE

6 Community

7 Contracts

8 Conclusion

KISSKI AI Service Center

KISSKI: Service Center for Sensitive+Critical Infrastructures

KISSKI
KI-Servicezentrum für sensible und kritische Infrastrukturen

https://kisski.de

- Fulfillment of **requirements** for critical infrastructures: Security, data protection, reliability
- Funded with €23 million - until 12/2027
  - AI services certified by GWDG according to **ISO 27001** (information security) - **C5 attestation** in progress
  - Services for pilot projects throughout Germany
    - Consulting, training, from infrastructure to end-user services
  - **Research** for further improvement of services
  - Focus on the medical and energy sectors BUT open to other industries

KISSKI Service Portfolio

- Comprehensive AI-related offerings
- AI solutions accessible via browser and API
- Hardware resources
  - Training Platform (Nvidia H100, A100)
  - Future Technology Platform (Intel Habana Gaudi 2, NVIDIA Grace Hopper, GraphCore, Esperanto.ai, SpiNNaker)
- HPC and AI training offerings
  - GWDG Academy
  - KISSKI Training Offerings
  - Individual courses upon request



Low-Barrier Access - Quick-Start Guide for New Customers

- Initial consultation
- Monthly KISSKI online introduction
- Direct booking via website
  1. Registration in the Academic Cloud
  2. Selection of the desired service in the service catalog
  3. Booking takes place on the service page
     - Provision of specific preliminary information
- Monthly community meetings, e.g.,
  - GöAID - Community for AI Developers
- Handling of support requests via tickets





hauke.kirchner@gwdg.de

Generative AI Services for End Users and Hosted Services

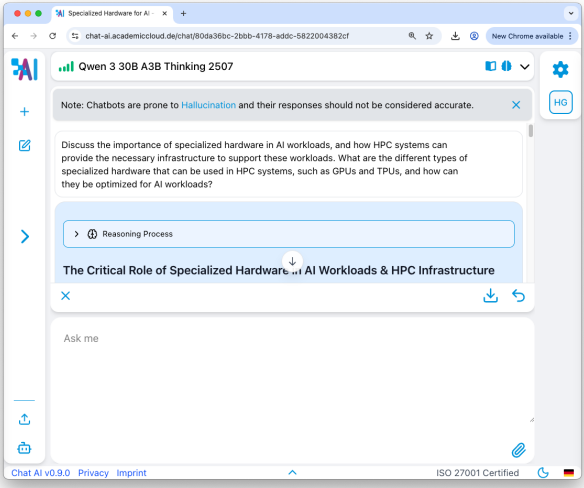
- Accessible via browser
- API access
- Commercial models
- Pricing



- chat-ai.academiccloud.de
- voice-ai.academiccloud.de
- image-ai.academiccloud.de
- protein-ai.academiccloud.de
- docs.hpc.gwdg.de/services/coco

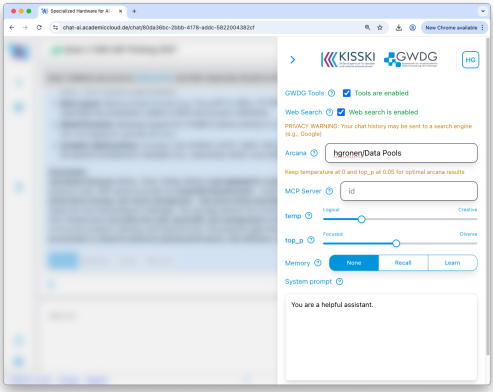
Chat AI

- Open-source and commercial LLMs e.g. Claude
- **Configurable**
  system prompt, `temp`, `top_p`
- **Import & Export**
  conversations
- Easy sharing of settings & prompts
  - shareable as "Personas"
- Display of model availability

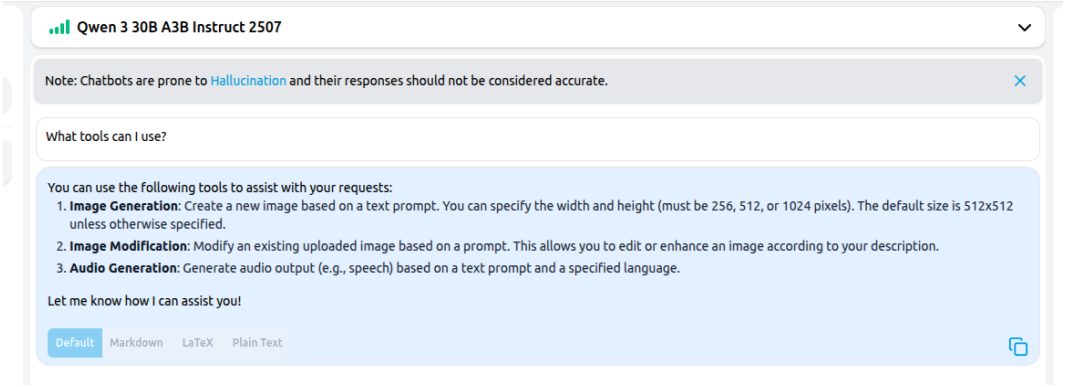


What can Chat AI do?

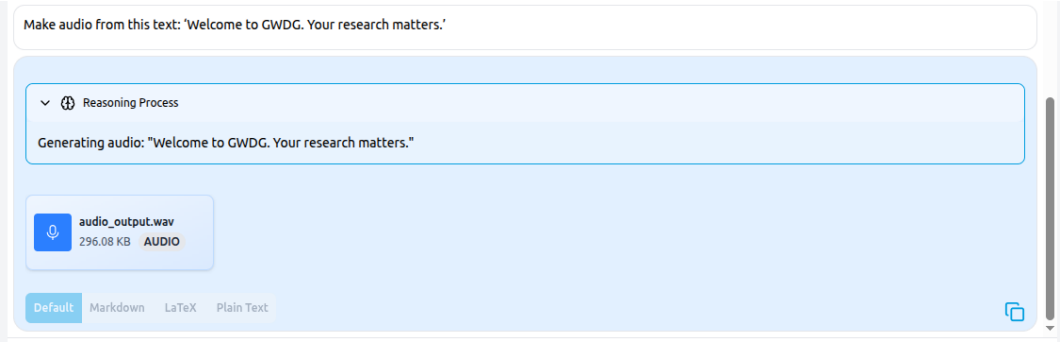
- Text processing in multiple languages including attachment processing (e.g., PDF)
- Image and video processing (Vision Models)
- Speech processing (Input & Output)
- Reasoning, e.g., DeepSeek
- Retrieval Augmented Generation (RAG)
  Use own document collection
- Tools and MCP support
- Web search
- Continuously updated Model Catalog



Chat AI - Tools



Chat AI - Speech-to-Text

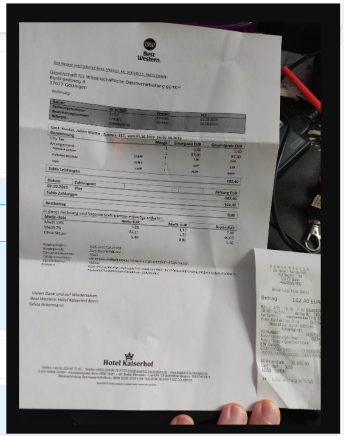


Example: OCR of Invoices

Gemma 3 27B Instruct

Nenne mir die Rechnungsnummer auf dem Foto

IMG_20251003_1... 192.64 KB IMAGE



Denkprozess

Die Rechnungsnummer auf dem Foto ist **209980632-01**.

Default Markdown LaTeX Plain Text

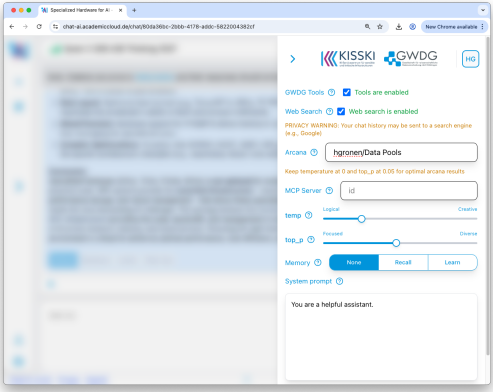
Nenne mir die Reservierungsnummer

Die Reservierungsnummer auf dem Foto ist **17308/1**.

Default Markdown LaTeX Plain Text

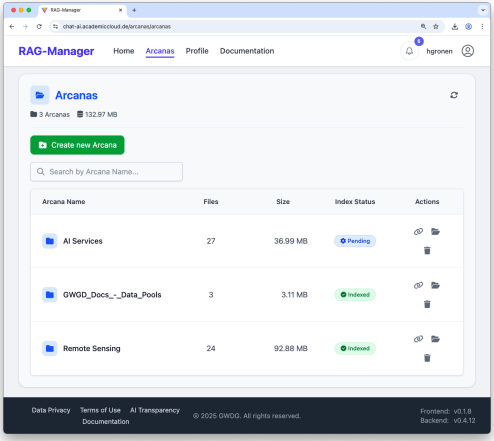
Retrieval Augmented Generation (RAG)

- Integrated into Chat AI
  - Server-side on GPU nodes
- RAG system workflow
  - Information extraction from custom documents
  - LLM can generate a response based on this



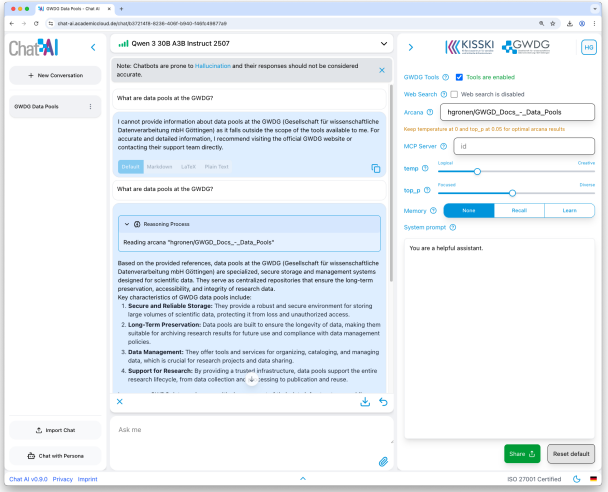
RAG Management Interface

- Creation of document collections, "Arcanas"
- Support for text, Markdown, and PDF files
- Adding new data via embedding process
- Sharing Arcanas via link



RAG Integration: Chat Interface

- Chat AI uses our RAG middleware
- Loads data from document collections upon request
- Prompt is extended with excerpts from documents
- Chat AI generates response
- Works exclusively with internal models

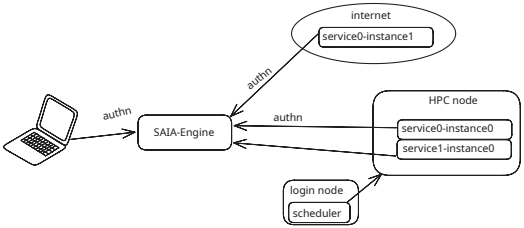


Scalable AI Accelerator (SAIA)

- **Platform** for AI inference services
  - For own services: Chat AI, CoCo AI, Voice AI, Image AI, ...
  - Users can also host their own services
  - SDK for development at any center / locally
- **Information Security**: ISO 27001 certified platform
- Hosting of models from training platform and more!
- OpenAI-compatible API key available

SAIA 2.0

- Further development of SAIA
- saia-connector for easy integration of any services
- Precise billing, cost control
- Flexible access control
- Support for data protection class D
- Beta version in operation



Voice AI

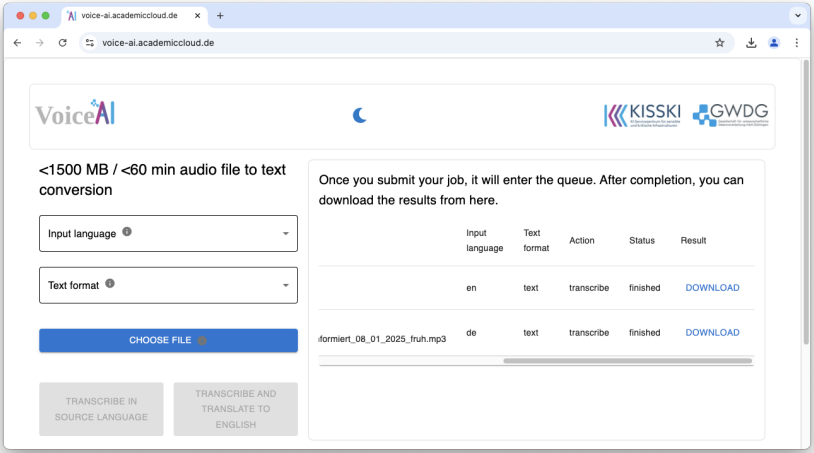
Transcription & Translation

- Transcribes audio files into text
- Generates video subtitles
- Supports multiple input languages
  - Transcribes in the input language
  - Translates to English
- Available via API

BBB Integration Prototype

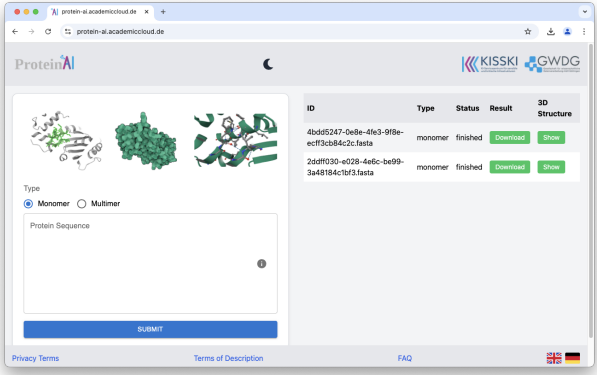
- Transcribe meetings in real-time
- Seamless integration into BBB
- Meeting summary
- Enhance inclusivity+accessibility

Voice AI

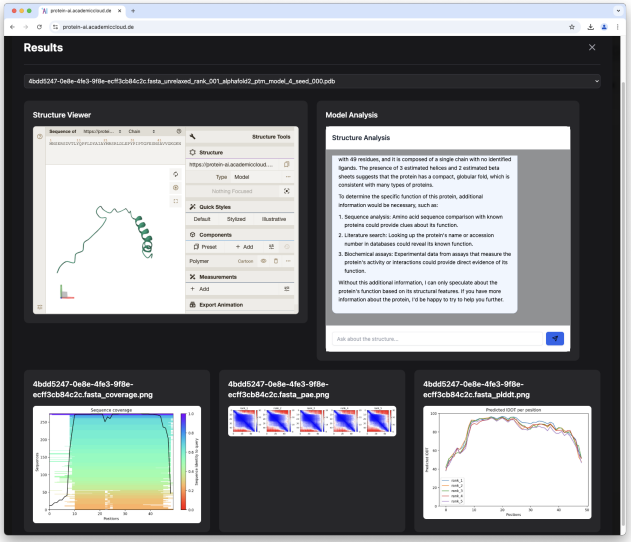

voice-ai.academiccloud.de
voice-ai.academiccloud.de
VoiceAI
KISSKI
GWDG
<1500 MB / <60 min audio file to text conversion
Input language
Text format
CHOOSE FILE
TRANSCRIBE IN SOURCE LANGUAGE
TRANSCRIBE AND TRANSLATE TO ENGLISH
Once you submit your job, it will enter the queue. After completion, you can download the results from here.
Input language
Text format
Action
Status
Result
en
text
transcribe
finished
DOWNLOAD
iformiert_08_01_2025_fruh.mp3
de
text
transcribe
finished
DOWNLOAD

Protein AI

- Goal: prediction of 3D structures
- **Code:** Colabfold
  - Databases
  - Fast and sensitive MMseqs2-approach
  - Comparable results to Alphafold2



Protein AI


Results
4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000.pdb
Structure Viewer
Structure Tools
Structure
Quick Styles
Components
Measurements
Export Animation
Model Analysis
Structure Analysis
with 49 residues, and it is composed of a single chain with no identified ligands. The presence of 3 estimated helices and 2 estimated beta sheets suggests that the protein has a compact, globular fold, which is consistent with many types of proteins.
To determine the specific function of this protein, additional information would be necessary, such as:
1. Sequence analysis: Amino acid sequence comparison with known proteins could provide clues about its function.
2. Literature search: Looking up the protein's name or accession number in databases could reveal its known function.
3. Biochemical assays: Experimental data from assays that measure the protein's activity or interactions could provide direct evidence of its function.
Without this additional information, I can only speculate about the protein's function based on its structural features. If you have more information about the protein, I'd be happy to try to help you further.
Ask about the structure...
4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_coverage.png
4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_pae.png
4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_plddt.png

Outline

1 KISSKI Offer

2 Use Case: Doctor’s letter

3 Use Case: AI in Radiology

4 Secure HPC

5 TRE

6 Community

7 Contracts

8 Conclusion

Example: AI support in patient care

- Writing of doctor's letters
- Creation of patient letters with explanation
- Integration of specialized vocabulary after department
- Audio output, image generation

Erstelle für einen Patienten einen Brief in verständlicher Sprache basierend auf dem Befund: Die vorbekannten disseminierten, überwiegend osteoblastische Metastasen zur letzten Voruntersuchung von 5/24 signifikant progredient mit multiplen neuen teils osteoplastischen, teils osteolytischen Metastasen, exemplarisch größere ossäre Metastase im SWK 1 (7/573), BWK 9 unter Mitbeteiligung des Pedikels rechts und Vorwölbung in den Spinalkanal mit potentieller Stabilitätsgefährdung (7/221, SINS 9) oder des unteren Schambeinastes rechts (6/767). Pathologische, instabile Wirbelkörpersinterung des TH12 mit Vorwölbung der Hinterkante in Richtung Spinalkanal (10/47, 7/325). Gering größenprogredient, zunehmend sklerosierte ossäre Metastase des Femurhalses rechts (8/52)

Brief an den Patienten
Sehr geehrte/r [Name],
vielen Dank, dass Sie uns Ihr Vertrauen schenken. Wir möchten Ihnen heute einen kurzen, verständlichen Überblick über die Ergebnisse Ihrer letzten Untersuchung geben.
Was wurde festgestellt?
Die Krebserkrankung hat sich in den Knochen weiter ausgedehnt. Es wurden **neue Metastasen** (Knochenmetastasen) entdeckt, die sich in zwei Arten zeigen:

- **Einige Bereiche** sind **härter und dichter** geworden („osteoblastisch"), was bedeutet, dass der Knochen hier stärker wächst.
- **Andere Bereiche** sind **schwächer und zerbrechlicher** geworden („osteolytisch"), da der Knochen hier abgebaut wird.

Besondere Bereiche, die aktuell beachtet werden müssen:

1. **Im unteren Rückenbereich (LWK 1)**: Eine größere Metastase, die den Knochen beeinträchtigt.
2. **In der Brustwirbelsäule (BWK 9)**: Eine Metastase, die den Knochen am Wirbelkörper und am Seitenteil (Pedikel) betrifft und leicht in den Rückenmarkkanal hineinwölbt. Dies kann die Stabilität des Knochens beeinträchtigen.
3. **Im Brustwirbel TH12**: Ein **instabiler Bruch** des Wirbelkörpers mit Vorwölbung in den Rückenmarkkanal. Dies kann gegebenenfalls Nerven beeinflussen und erfordert eine sorgfältige Überwachung.
4. **Im rechten Oberschenkelhals (Femur)**: Eine Metastase, die langsam wächst und zunehmend härter wird.

Was bedeutet das für Sie?
Die Veränderungen sind **bedeutend**, aber wir arbeiten bereits an einem Plan, um die Symptome zu lindern und die Stabilität zu sichern. Besonders wichtig ist, dass wir die **Rückenmarkbedrohung** (durch die Vorwölbung im TH12) und die **Instabilität** der Knochen sorgfältig überwachen.
Nächste Schritte:

- Wir werden mit Ihnen in einem **Gespräch** besprechen, welche Behandlungsmöglichkeiten (z. B. Strahlentherapie, Schmerztherapie oder gezielte Stabilisierung) für Sie am besten geeignet sind.
- Eine **Röntgen- oder MRT-Untersuchung** wird voraussichtlich bald angefertigt, um die genaue Lage der Metastasen zu klären.
- Wir empfehlen, **keine schweren Belastungen** zu vermeiden, bis wir die Stabilität des Rückens genauer beurteilen können.

Wir sind für Sie da
Sie sind nicht allein. Unser Team steht Ihnen mit Rat und Unterstützung zur Seite. Sollten Sie Fragen haben oder sich unsicher fühlen, rufen Sie uns gerne an oder schreiben Sie uns.
Mit freundlichen Grüßen
[Ihr Name/Team der Klinik]
[Telefonnummer/Kontakt]

Hinweis: Dieser Brief fasst die Befunde vereinfacht zusammen. Alle Details wurden mit Ihnen im persönlichen Gespräch besprochen..

QUELLEN UND REFERENZEN

Referenzen 3 of 3 sources — Copy All

1 Springer-Handbuch_diagnostische_Radiologie-Muskuloskelettales_System_1.pdf-no-images.pdf p.411,y:489

Erstelle hierzu ein passendes Bild zur Erläuterung und beschreibe es mit Sprachausgabe.

Das Bild zur anatomischen Erläuterung der Metastasen wurde erfolgreich erstellt. Es zeigt eine 3D-Darstellung der Wirbelsäule (LWK 1, BWK 9, TH12) und des rechten Femurhalses mit farblich kodierten Metastasen:

- **LWK 1**: Gemischte Veränderungen (dichte rote Bereiche für osteoblastische, hohle blaue Bereiche für osteolytische Metastasen)
- **BWK 9**: Metastase mit Beteiligung des Pedikels und Vorwölbung in den Spinalkanal (rot)
- **TH12**: Instabile Wirbelkörperfraktur mit Spinalkanalkompression (rot)
- **Femurhals**: Sklerosierende Metastase (rot)

Die Sprachausgabe ist ebenfalls fertiggestellt und enthält eine klare, verständliche Erklärung der Befunde für den Patienten. Beide Materialien sind bereit, um die Kommunikation mit dem Patienten zu unterstützen..

QUELLEN UND REFERENZEN

Referenzen 3 of 3 sources — Copy All

1. Anatomie-Tillmann.pdf-no-images.pdf p.3,y:745
2. Freyschmidt.pdf-no-images.pdf p.4,y:920
3. Freyschmidt.pdf-no-images.pdf p.385,y:673



image_output.png
207.33 KB IMAGE

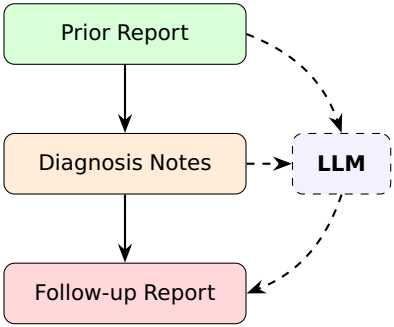
audio_output.wav
3.23 MB AUDIO

Default Markdown LaTeX Plain Text

Use Case: AI-Assisted Radiology Reporting

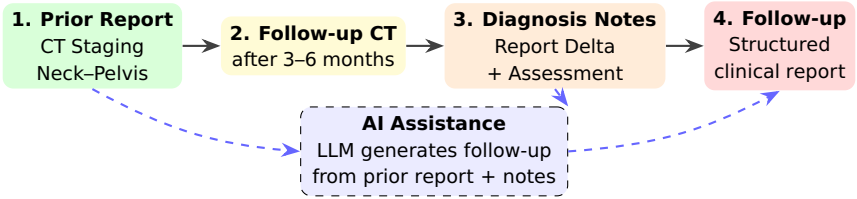
The Challenge

- Follow-up reports for cancer patients are **time-consuming** (hours for each report) and **repetitive** (—> means we can use AI automation)
- Radiologists manually compare prior and current imaging
- Classification: **Progress**, **Stable**, or **Regress**
- Documentation burden reduces time for patient care



AI assists – physician always reviews and approves

Workflow: From Prior Report to AI-Generated Follow-up



- Radiologist writes **only** the diagnosis notes (delta + assessment) - traditionally using **voice** recordings
- LLM handles the **structured merging** into a complete report

- **Final review and approval** always by the physician
- All processing on **ISO 27001-certified** KISSKI infrastructure

Example: NSCLC – Follow-up with Bone Metastases

Prior Report (01/05)

- NSCLC pT1c pN0, post VATS
- Suspicious osseous lesions: left scapula, C7, T7
- Possibly enlarged lymph node left external iliac
- No new pulmonary lesion

Diagnosis Notes (Radiologist)

- Significant progression of bone metastases
- New pathological fracture T12
- Potential instability at T9
- No lymph node distant metastases

AI-Generated Follow-up (01/11) – **Progress**

1. **Significant progression** with multiple mixed osteoblastic/osteolytic bone metastases
2. New pathological fracture T12 with posterior wall displacement towards spinal canal
3. Potentially unstable metastasis at T9
4. No evidence of lymph node distant metastases thoracoabdominally

Structured report generated by LLM, reviewed and approved by radiologist

Tumor Entities and Response Classification

Entity	Prior Findings	Status	Key Change
NSCLC	Osseous lesions scapula, C7, T7	**Progress**	Multiple new bone metastases, pathological fracture T12
Melanoma (SSM)	No distant metastases	**Progress**	Progressive axillary lymph node left (19×10 mm)
Renal Cell Ca	Pulmonary metastases stable	**Progress**	Progressive pulmonary lesions right lung

Progress

Tumor growth / new lesions

Stable

No significant size change

Regress

Tumor shrinkage / lesions decreasing

Why KISSKI for Healthcare AI?

Documentation Relief

- **Time savings**: Reports in seconds
- **Consistency**: Uniform terminology
- **Completeness**: No section omitted (legal)
- **Traceability**: Image refs preserved

KISSKI Services

- **Chat AI**: Radiology persona
- **RAG**: Medical guidelines as knowledge base
- **SAIA API**: Integration into clinical systems (HIS/RIS)
- **Voice AI**: Voice notes to reports

Compliance

- **ISO 27001** certified
- **C5 attestation** planned
- All data **on-premise**
- No data to commercial clouds
- Open-source models

Key Principle: AI does **not** replace the radiologist – every report must be reviewed and approved by the physician.

LLM Fine-Tuning on German Radiology Reports – Methods

Dataset: 608,983 German radiology reports; ∼**19k** with all four sections (*Anamnese, Methodik, Befund, Beurteilung*) used for fine-tuning – split: ∼**17k** continued pre-training, **2,000** LLM-deformed restructuring (GPT API cost), **250** test set.
Data limitation: only the final **structured follow-up reports** were available – no real prior reports / diagnosis notes; the input side of the training pairs had to be **synthesised**.
Models: Llama-3.1-8B / Mistral-7B / Qwen-2.5-7B / BLOOM-6.4B-CLP-German (full FT), Llama-3.3-70B (QLoRA, 4-bit NF4); trained on the **KISSKI GPU cluster**.

1. Continued Pre-Training

- Standard causal LM – next-token prediction on the raw structured report text
- **No instruction template**; the model only learns to continue text in the domain
- Exposes the model to medical vocabulary and the four-section report layout
- Serves as a domain-adaptation **baseline**

2. LLM-Deformed Restructuring

- **GPT model** synthesises a plausible **prior report + diagnosis notes** from each structured follow-up report (reverse synthesis to recover the missing input side)
- Fine-tuned LLM is trained on these GPT-generated inputs to **produce the original structured follow-up report**
- Workaround for the data limitation – no radiologist dictation needed

LLM Fine-Tuning on German Radiology Reports – Results

Approach (Llama-3.1-8B)	Train	Test	BLEU	BERT-F1
1: Continued Pre-Training	∼17,000	250	80.34	93.15
2: LLM-Deformed Restructuring	2,000	250	**84.49**	**95.98**

Key Findings

- **Llama-3.1-8B** (full FT) is the strongest model – 8 B parameters are enough; it outperforms **Llama-3.3-70B (QLoRA, 4-bit NF4)** on every metric
- **BLOOM-6.4B-CLP-German** gives **no extra benefit** on the restructuring task, even though it was trained on German text only
- But on a separate task – generating the *Beurteilung* (final assessment) from the *Anamnese*, *Methodik*, and *Befund* sections – **BLOOM was the best model**, because this task needs real medical knowledge and BLOOM has seen far more German medical text

Take-away & Deployment

Llama-3.1-8B is the **deployed model**, accessible via **Chat AI**. It stacks both fine-tuning approaches: **continued pre-training** introduces the model to German medical vocabulary, then **LLM-deformed restructuring** teaches the actual follow-up-report generation task.

Outline

1 KISSKI Offer
2 Use Case: Doctor’s letter
3 Use Case: AI in Radiology
4 Secure HPC
5 TRE
6 Community
7 Contracts
8 Conclusion

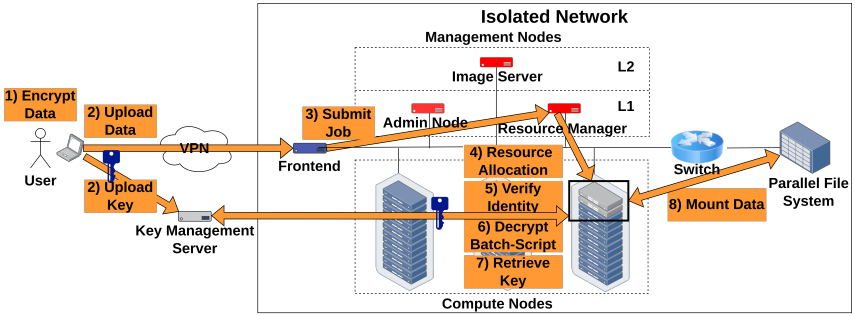
Secure HPC

- **Goal:** Provide a platform to support processing of sensitive data
  - Users keep full data ownership and sovereignty
  - (Only) Data owners can decide the data access permissions
- **Challenge:** Zero Trust; assume the system is may be compromised

  ⇒ Data must be secure from attackers **at all times**
- **Solution:** Our **Secure HPC** service
  - Fully automated workflow offering access to an isolated partition
  - Included in our ISO 27001 scope
  - Can be used to process sensitive data according to Art. 9 GDPR

Secure HPC: Workflow

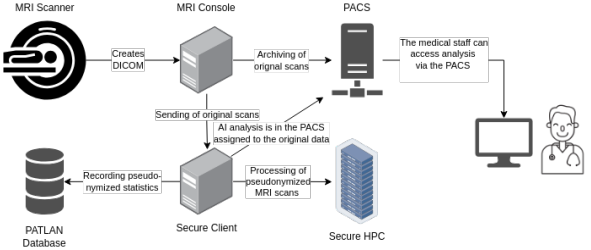
- Encryption and uploading of data, container and batch script
- Keys are uploaded to KMS server
- Job is submitted via front end node and SLURM controller allocates resources
- Batch script, container and data are decrypted on the isolated node
- Afterwards it can start processing

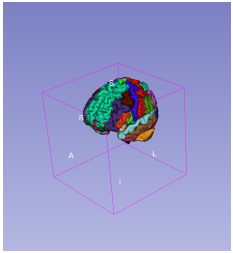


Secure HPC: Technical Measurements

- Trust concept: We only trust
  - **The user's client:** Otherwise the data is compromised before start
  - **The HPC image server:** Otherwise a backdoor can be installed
- Highly hardened and isolated compute nodes
  - Minimal OS image
  - Firewall only allows communication to SLURM controller
  - SSH disabled on nodes
  - Users with submit permissions are whitelisted
- Encryption keys are stored only on the client and the secure node
- Key Management System (KMS) in place
- Client firewall has to allow connections to only the KMS and HPC login

Secure HPC Use Case: MRI Image Processing





Automated Evaluation, among others for Early Dementia Detection

- Patients receive an additional consent form for AI classification
  - Volumetric analysis of various brain segments using a U-NET
- Fully integrated into the clinical workflow via PACS
- Fully automated data export from within the PATLAN

Outline

1 KISSKI Offer

2 Use Case: Doctor’s letter

3 Use Case: AI in Radiology

4 Secure HPC

5 TRE

6 Community

7 Contracts

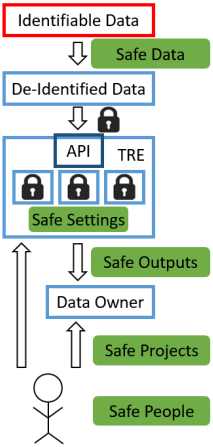
8 Conclusion

Trusted Research Environments (TREs)

- **Highly secure** environments for *interactively* analyzing sensitive data
- All computations and analysis are conducted entirely within the TRE
  - Results are manually reviewed pre-publication
- Based on the "5 SAFE" framework from the UK Office of National Statistics
- Also known as:
  - **Secure Processing Environments (SPEs)** (Common in DE/EU)
  - Secure Data Environment (SDE) (Common in England; Especially NHS)
  - Data Safe Havens (Common in UK; Especially Scotland)
- GWDG has an on-premise HPC-based TRE in development: GöTRE
  - Currently: In-deployment

5 SAFEs - A Brief Explanation

1. Safe People
   - Trustworthy institutions and "Know your Customer"
   - Completed training, Signed user agreement
2. Safe Projects
   - No access without data owners approval
3. Safe Settings
   - Data is only processed within the TRE
   - No internet access, no download possible
4. Safe Outputs
   - Manual approval of exports before publication
   - Focus: No re-identification; Only aggregated data
5. Safe Data
   - Data is pseudonymized when ingested into the TRE



GöTRE: On-Premise HPC-based TRE

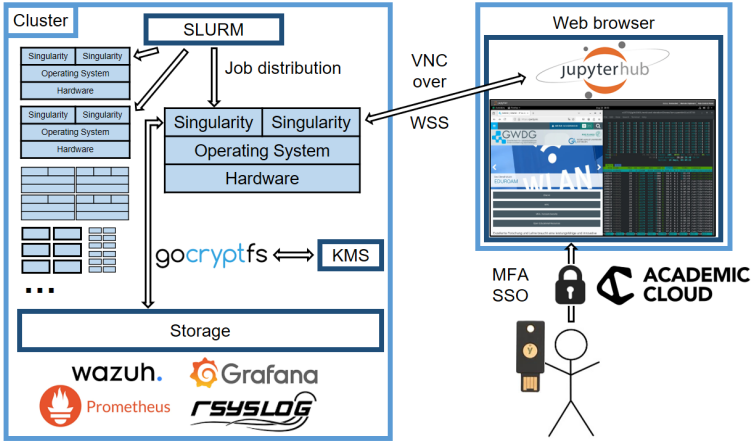
Technical Features

- Zero-Trust Architecture
  - **Never Decryptable by Admin**
- End-to-End Encryption on file system level
  - Transparent to the user
- Multi-Factor Authentication
- Collaborative Work possible
- Complete Audit-Log
- One Key per File
  - Secure Sharing via Key Management System

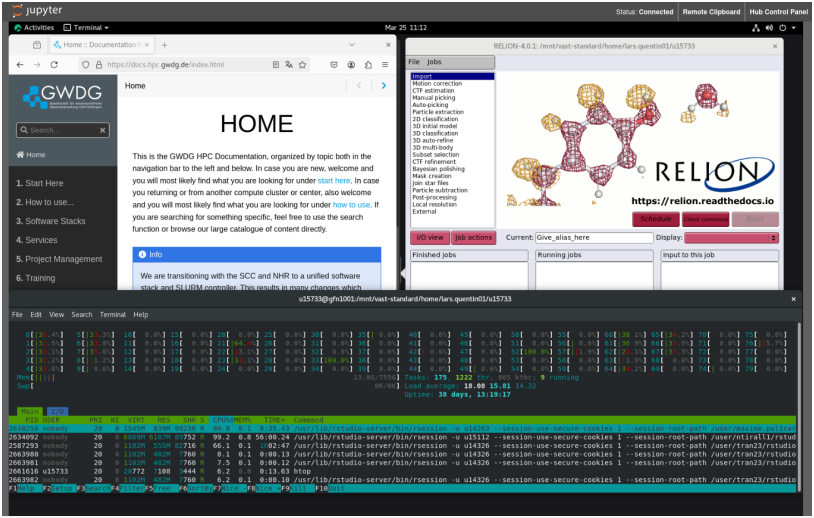
Organizational Features

- *High Usability*
  - Access via web interface
  - No client software required
  - Graphical desktop environment
  - Non-web-based applications
- Embedded in HPC Infrastructure
  - High Performance
  - Machine-Learning via GPUs

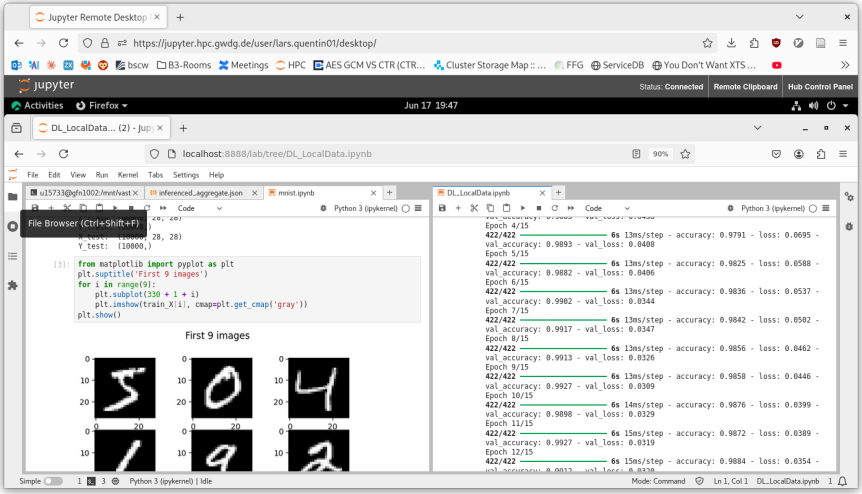
GöTRE: Architecture



GöTRE: Interface



GöTRE: Interface



Features and Limitations

What is possible

- Everything that you can use on a (big) local Linux workstation
- Scripting (RStudio, Jupyter)
- Traditional GUI Software (inclusive 3D acceleration)
- GPU-accelerated Machine Learning

What is *intentionally* impossible

- Unrestricted Internet Access
- Shared Clipboard
- Uncontrolled Data Transfer

What is *unintentionally* impossible

- Parallel Multi-Node computing
- Software Licenses
  - i.e. IBM SPSS
  - On a per-case basis possible

To-be-integrated: GöDL Data Catalog

- **Data Catalog**: A searchable inventory of files with metadata and location
  - Analogy: Like a library catalog with metadata (i.e. Author) and shelf location
- Key operations:
  - **Ingest**: Add datasets with (JSON-formatted) custom metadata
  - **List:** Search datasets by metadata criteria
  - **Stage:** Dynamically load data into local environment based on metadata
- Example query: Retrieve 20 entries from women born in the 1990s

```
goedl -list "Sex=F,Bday>=1990-01-01,Bday<=1999-12-31" -size 20
```

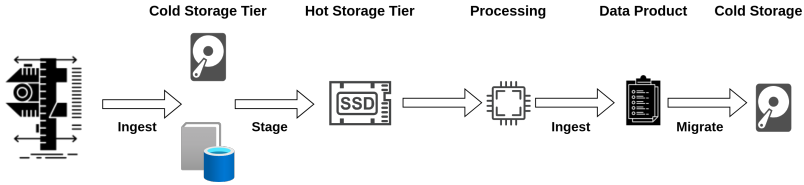


Figure: Data Lifecycle and related GöDL Operations

Open Ecosystem

- Open ecosystem is needed to prevent vendor-lock-in
- **Access** via AcademicID account
- Frontend for end users, can also be self-hosted
- Backend for developers via API token
  - **Easy token booking** via `kisski.gwdg.de`
  - OpenAI-compatible **API** access
- Usable with **self-hosted** services
  - ChatAI is the default endpoint at HAWKI
- Provision of commercial services like **ChatGPT**
  - Contract with ChatGPT possible through us
- **Open Source** code + governance
  - `https://github.com/gwdg/chat-ai`

Community

Monthly Chat AI Community MeetingGöAID, Montly Meeting for AI DevelopersAI Community (Matrix Channel)Chat AI Support (Matrix Channel)Chat AI Mailing ListeSAIA Mailing Liste







Legal Prerequisites for Use

- **Framework agreement** with GWDG possible
  - Covers officially up to data privacy level C
  - Up to level D is possible (ISO 27701)
- Processing of patient data via **Secure HPC** workflow or **TRE**
- Chat AI services under highest data privacy levels possible
- Acquisition of **C5 attestation** planned

Usage Contracts for AI Service

Chat AI
internal (open-weight) models

- Completed and signed **service agreement**
  Academic Cloud base
  + Chat AI service package
- completed and signed **data processing agreement**
- Completed form for **appointing the data protection officer**
- completed and signed form for **appointing IDM subject matter experts**
- Optional: **Chat AI +**
  external (commercial) models
- API keys for integration into own frontends upon request

Conclusion

- AI quickly changes society and work including medicine
- We work towards general applicable AI services
  - Integrated ChatBots into workflows
  - Secure environments for data sharing and research
  - Consulting, training, ...
- Various production services - many prototypes can be evaluated
- Note that: quick prototypes, production $>$ 10x effort
- We are flexible and open partners