LUHKI-HPC

LUHKI-HPC is an AI-powered assistant designed to support users of the LUIS High Performance Computing (HPC) cluster. It provides an interactive interface to the official cluster documentation, enabling both current and prospective users to ask questions about system usage, software and hardware environments, recommended practices, and selected aspects of their personal cluster environment.

Notice: The system is currently in a **Test Phase** and under active development.

Please ensure that you have read and accepted Terms of Use and Privacy Policy before using LUHKI-HPC. Follow the link for the German version.

System Overview

LUHKI-HPC is based on a Retrieval-Augmented Generation (RAG) architecture. Before generating a response, the system retrieves relevant content from the official LUIS documentation and uses it as context.

All AI models are open-weight models hosted locally on LUH cluster infrastructure. Model execution is performed on compute nodes within the cluster environment, typically without external network access. User input and generated responses are processed entirely within LUH infrastructure and are not transmitted to external third-party services. Any APIs used are strictly internal to the LUH cluster infrastructure.

The underlying RAG knowledge base is synchronized with the official cluster documentation on a daily basis. Updates to the documentation may therefore be reflected in LUHKI-HPC with a delay of up to 24 hours.

LUHKI-HPC currently provides two modes of operation:

Guest Mode

Guest mode is available from the official cluster documentation page and does not require user authentication.

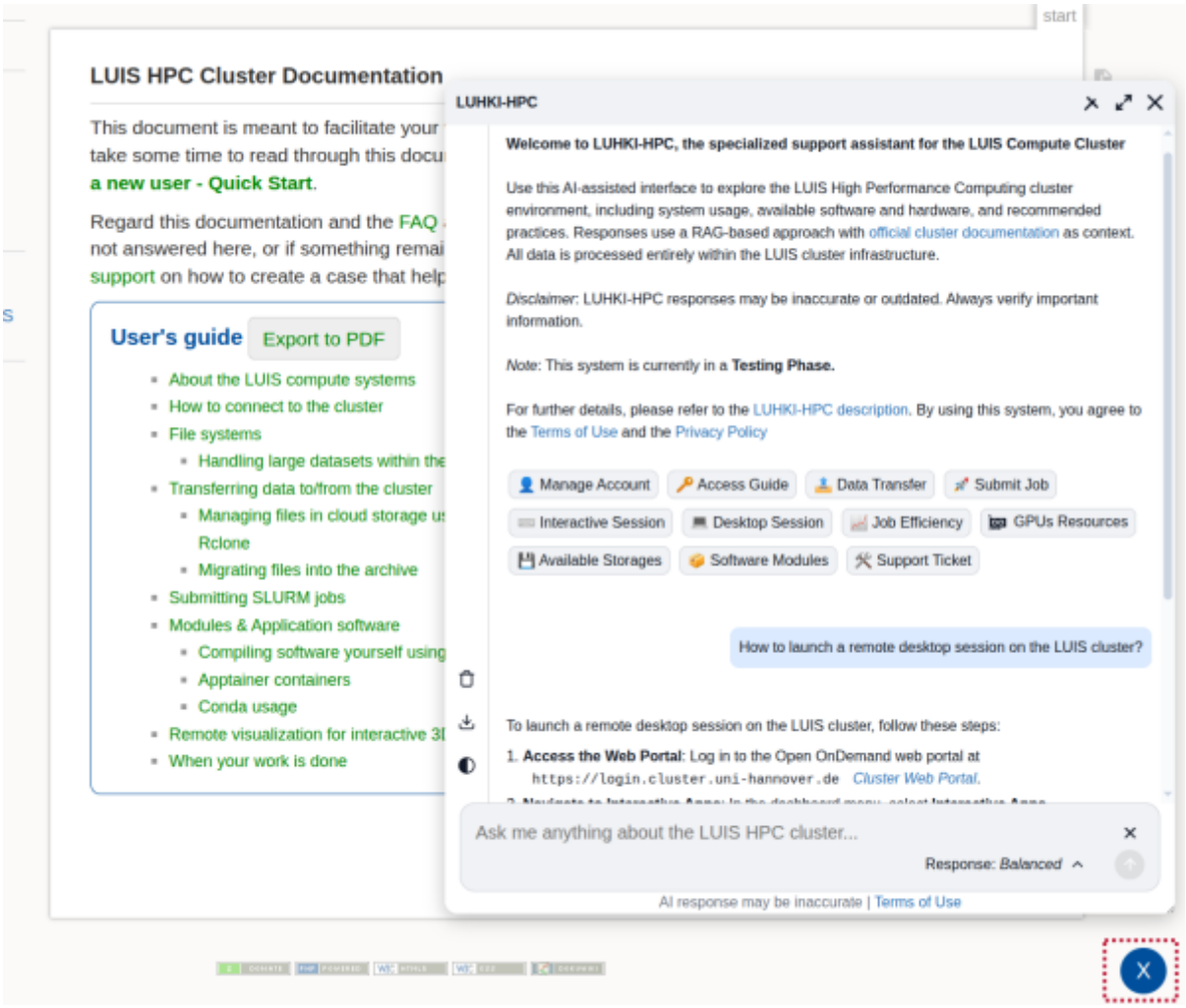

Fig. 1: LUHKI-HPC on cluster docs site

In this mode:

- LUHKI-HPC retrieves cluster-specific information primarily from the official LUIS cluster documentation provided through the RAG knowledge base system.
- The assistant has no access to the user's cluster account, files, environment, or running jobs.
- Conversation history is stored locally in the browser using *sessionStorage* to prevent accidental message loss during browser page reloads.
- Stored conversation data is automatically removed once the browser tab is closed.
- Only a single temporary chat session is available.

Authenticated Mode

Access to the authenticated mode is provided through the Open OnDemand platform and requires a valid cluster account.

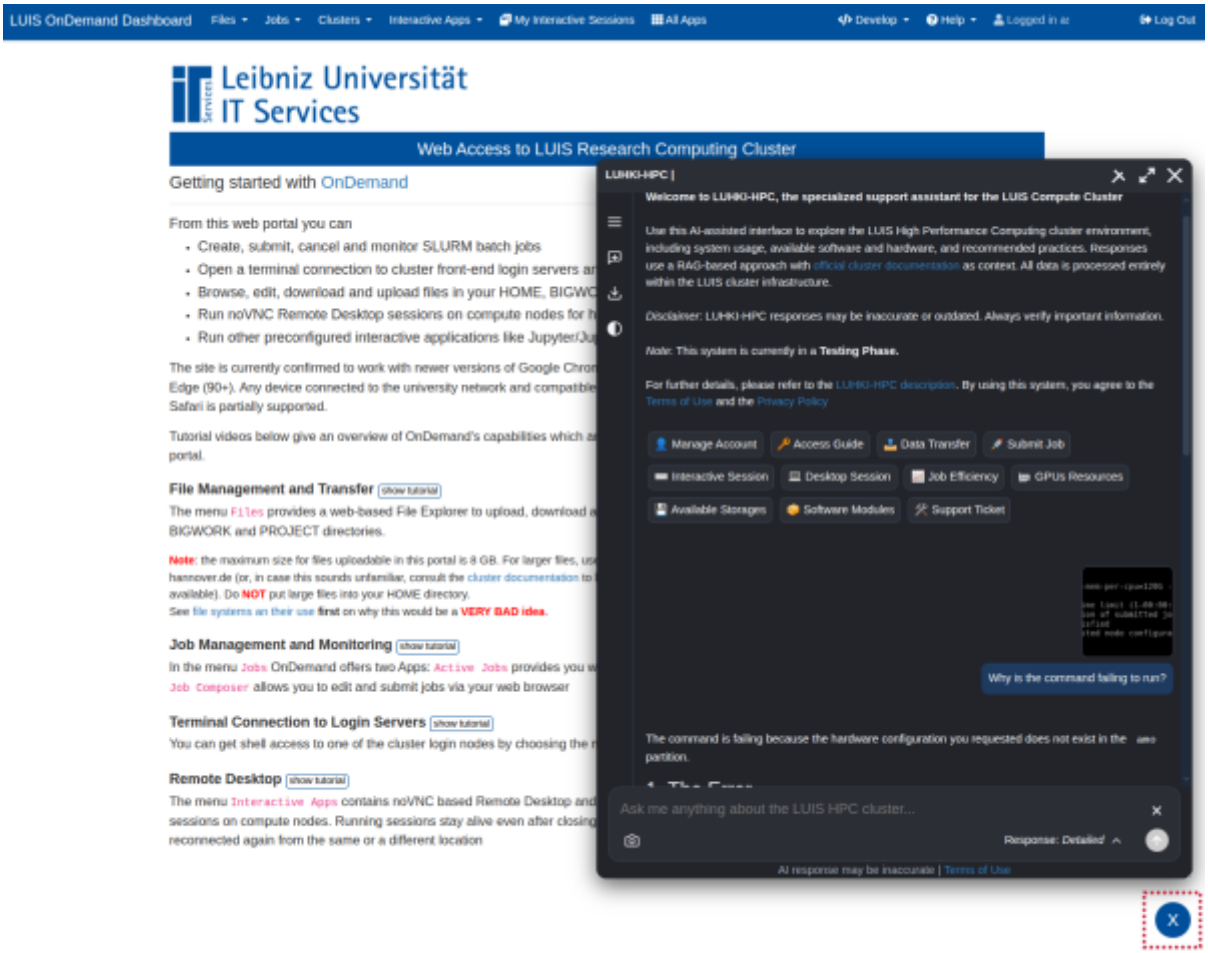

Fig. 2: LUHKI-HPC available on Open OnDemand

In this mode:

- LUHKI-HPC has limited access to the authenticated user's cluster environment.
- In addition to the RAG knowledge base, LUHKI-HPC may execute selected shell and Slurm commands within the user's cluster account context to provide environment-aware assistance.
- Users can upload images (JPEG/PNG/WebP) directly into the chat interface to provide visual context, subject to a maximum size limit of 8MB per file.
- Chat sessions and conversation history are stored persistently in the user's home directory on the cluster using a local SQLite database.
- Multiple chat sessions are supported.
- Previously created chat sessions can be reopened, renamed, or removed through the sidebar interface.

User Interface Features

Users can select between three AI response modes via the “Model Response” menu:

- **Brief Answers**: Optimized for low latency and concise answers.
- **Balanced Guidance**: Default setting, providing a balance between response speed and explanatory detail.
- **Detailed Solutions**: Intended for in-depth explanations with slower but more comprehensive responses.

Additional interface features include:

- Persistent multi-session chat management in authenticated mode.
- A sidebar interface for tools and managing chat sessions.
- Conversation export in Markdown format for documentation or sharing.
- Options to clear conversation history and reset the input field.
- Support for both default and expanded chat display modes.
- A switch for toggling between light and dark interface modes.
- Automatic formatting of code snippets and command-line examples.
- One-click copying of code blocks and terminal commands to the clipboard.

Limitations and Notes

LUHKI-HPC responses may be inaccurate, incomplete, or outdated. Users should always verify important technical information before applying them in practice.

The system is currently considered **experimental** and remains under active development.

LUHKI-HPC is restricted to the LUH network and other explicitly authorized environments. Access from external sources may therefore be limited or blocked.

Terms of Use

Service Scope and Applicable Policies

LUHKI-HPC is provided as part of the LUIS “Scientific Computing” services and operates on shared LUH cluster resources. Use of the service is subject to the general policies governing LUIS IT systems and services as well as the official LUH cluster usage guidelines.

Users are required to comply with these policies when using LUHKI-HPC.

Service Availability

LUHKI-HPC is available under the same general conditions as the other resources of the LUH computing cluster. Continuous or uninterrupted operation is not guaranteed.

Nature of Service and Disclaimer

LUHKI-HPC provides AI-generated responses based on the official LUH cluster documentation through a Retrieval-Augmented Generation (RAG) approach. In authenticated mode, the system may additionally use selected information retrieved from the user's cluster environment.

All responses are generated automatically and may contain inaccuracies, omissions, or outdated information. The information provided by LUHKI-HPC does not constitute official, authoritative, or binding guidance.

Users are solely responsible for verifying all generated outputs before applying them in practice, particularly in production, security-relevant, or resource-critical environments.

Privacy Policy

Reference to Existing Policies

Use of LUHKI-HPC is subject to the general data protection and privacy policies applicable to LUIS IT services and the LUH cluster infrastructure. These policies define the overarching legal and organizational framework governing data processing and system usage.

Data Processing

User queries and generated responses are processed by locally hosted Large Language Models operating within the LUH cluster infrastructure. No data is transmitted to external third-party providers or external AI services.

Conversation Storage

In guest mode, conversation data is stored temporarily in the user's browser *sessionStorage* solely to prevent accidental data loss during browser page reloads. This data remains entirely on the user's device and is automatically removed once the browser tab is closed.

In authenticated mode, chat sessions and conversation history are stored persistently in the user's home directory on the cluster using a local SQLite database in order to support multi-session chat functionality.

Server-side Logging

For operational, maintenance, and security purposes, server-side logs are maintained. These logs may include technical metadata such as:

- timestamps
- system events
- selected model configuration
- error and performance information

The content of user queries and generated responses is not stored persistently in server-side operational logs.