

HELMHOLTZ

nfdi  
Nationale  
Forschungsdaten  
Infrastruktur

GEOMAR



# DataXplorers Hackathon

Question answering about Research Data Management

- Question answering about Research Data Management
  - Provided by Dr. Auriol Degbelo and Dr.-Ing. Christin Henzen
  - Develop a web-based Q&A-Tool to assist researchers in finding answers related to research data management
  - Should be developed in Python



**Marcus Krüger**

❖ Data Manager at GEOMAR



**Everardo Gonzalez**

❖ Data Scientist at GEOMAR



**Alexander Pilz**

❖ Intern at GEOMAR



Researcher

- I want to retrieve information about RDIs in general or a specific RDI like [PANGAEA](#).



Provider

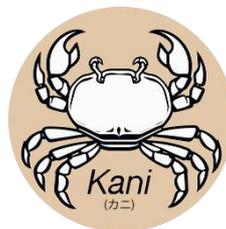
- I run a RDI and would like to enable my users to inform themselves quickly and easily about my RDI. I also do not want to create loads of new content.



Researcher & Provider

- We would like to perform preliminary checks on data that is to be uploaded to our RDI for compatibility with a data policy.

- How can we get access to *performant* and *non-proprietary* LLMs?
  - Host model locally, via Kani for example
    - **Many LLMs available**
    - **Requires a lot of computing capacity**
  - Host model via Institution
    - [Blablador](#) hosted by Helmholtz AI at the FZ Jülich
      - Accessible by members of the Helmholtz community
      - Hosts a selection of LLMs (Mistral, Yi, openChat etc.)



- Application developed with flask
- Provides API and simple web frontend
- Supports
  - Website ingestion
  - Inquiries about RDIs
  - Basic preliminary tests of .csv files
- Easily configurable via file
- Provides openAPI documentation



# Flask

## PANGAEA Chatbot

Interactiontool for Data Publishers of Earth & Environmental Science Data

### Introduction

#### Question answering about Research Data Management

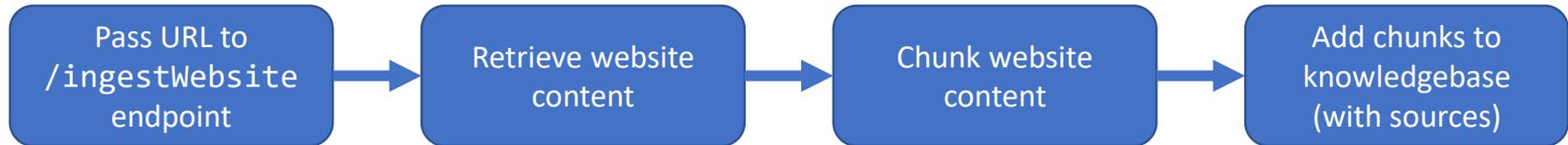
This challenge was provided by Dr. Aurilio Degbello and Dr.-Ing. Christin Henzen from the TU Dresden. Researchers face various challenges at each of the phases of the research data management lifecycle (data creation, data processing, data analysis, archiving, sharing and re-use). Depending on the nature of the challenge, they may turn to online platforms (e.g., Stackoverflow, GitHub, Quora, ...) or peers to seek answers. While online platforms and existing search engines may provide help at times, they fall short because either they are (i) general purpose and/or (ii) they return websites that researchers need to go through manually but not actual answers. A web-based system where they can ask their questions related to research data management easily and receive meaningful answers to these could yield productivity gains while also contributing to knowledge exchange among peers.

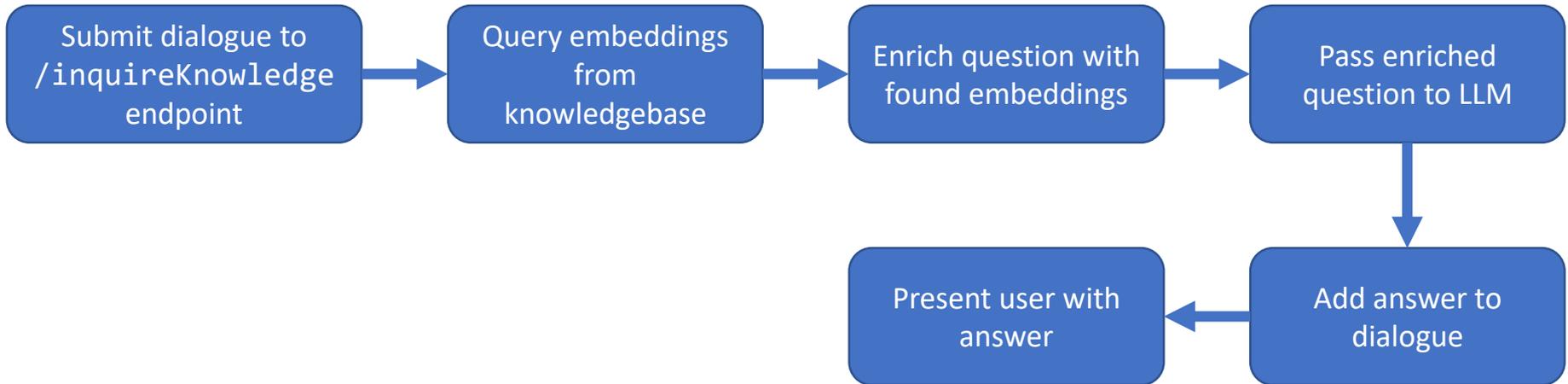
#### Goal

The goal of the Hackathon is to build a web-based question-answering system to assist researchers in finding answers related to research data management. Researchers ask their questions to the system in natural language and receive answers to their questions in natural language as well.

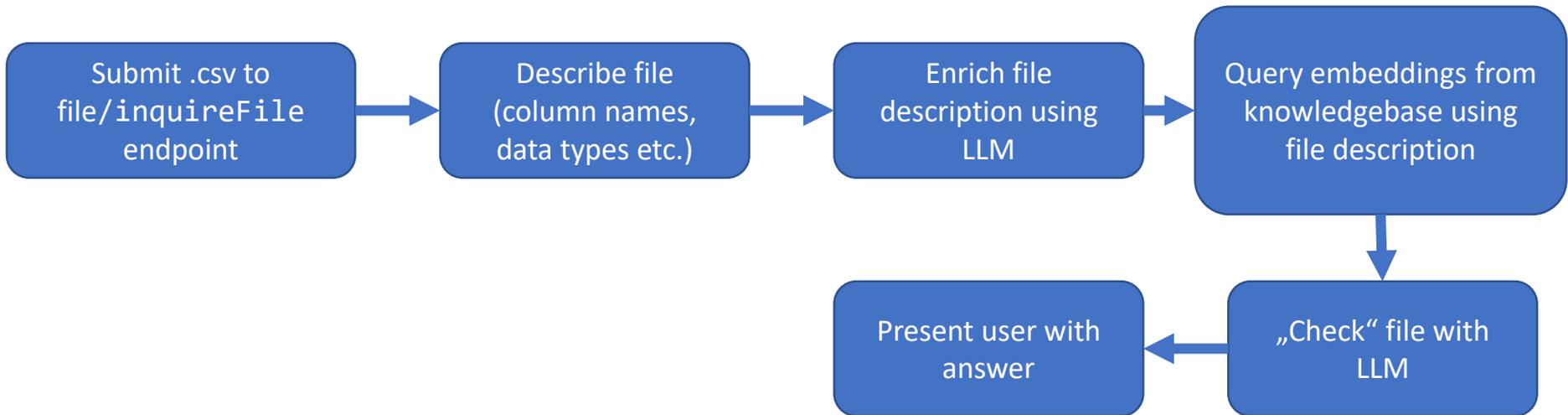
#### Data

A series of questions and answers related to RDM following the format Q/A/S (Question/Answer/Source). The dataset will be provided as a text file.





# Preliminary tests of .csv files



- Enhanced preliminary checks
  - Support for more formats
  - Run checks with [data curation jupyter notebooks](#)
  - Solve possible issues via function calling
- Additional administrative endpoints
  - Deletion of embeddings
  - Options to dynamically change configuration
- Compare different LLMs

HELMHOLTZ

nfdi  
Nationale  
Forschungsdaten  
Infrastruktur

GEOMAR



Thank you for your attention!

Any questions or remarks?

- How does such an application scale (if employed on PANGEA or NFDI4Earth level)?
- Could blablador be opened up for NFDI4Earth members?
  - Could an Base4NFDI service akin to Blablador be instituted?
- Could such a software also be used for searching datasets (Search Engine, rather than Q&A Tool)?
  - How to deal with a spatial component in inquiry?
- How could uploaded files really be checked against a data curation checklist?
- Could multiple languages be supported?

- Improvements to the Frontend and the UX
  - Alternatively, just deliver API
- Compare LLMs
- Preliminary file checking
  - Use data curation guidelines in natural language too (instead of Jupyter Notebooks)
    - Alternatively pass structured data descriptions (.json, .yaml etc.)
    - Employ code generating AI
  - Run actual checks on files
    - Return Report with recommendations to user
    - Perform necessary changes to files automatically