
HELMHOLTZ
nfdi
Nationale
Forschungsdaten
Infrastruktur
GEOMAR
DataXplorers Hackathon
Question answering about Research Data Management

Challenge

- 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

Team

Marcus Krüger

- Data Manager at GEOMAR

Everardo Gonzalez

- Data Scientist at GEOMAR

Alexander Pilz

- Intern at GEOMAR

User-Stories

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.

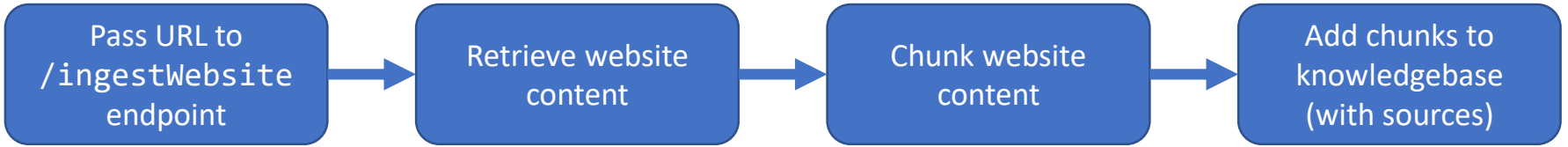
Blablador

- 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.)

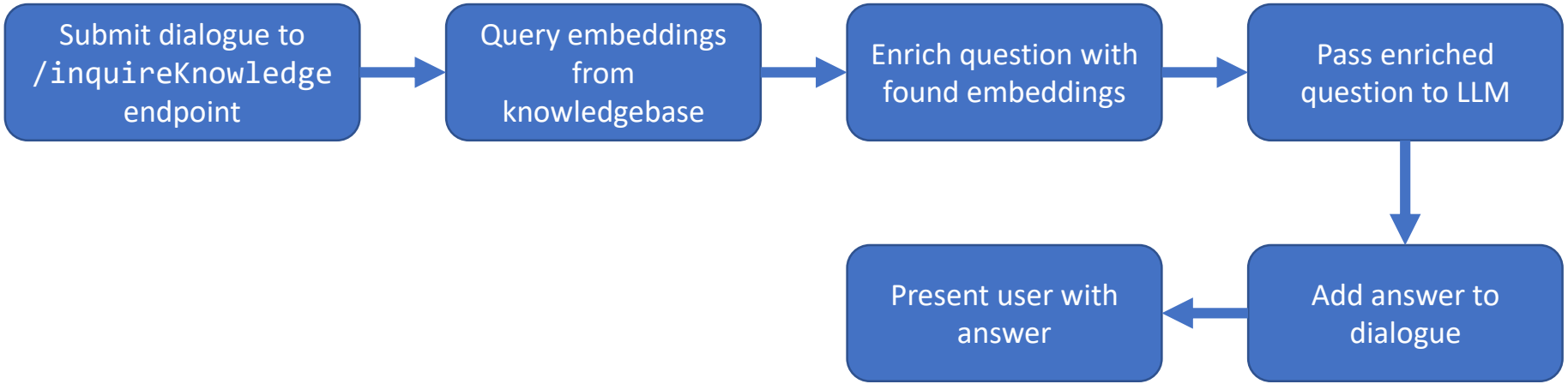
Application

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

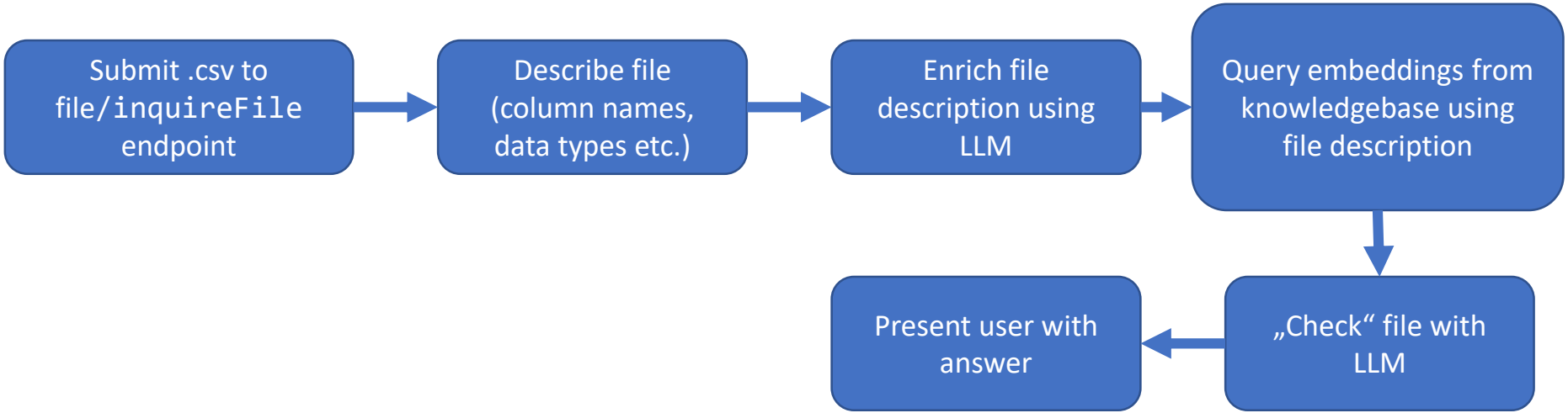
Website ingestion



Inquieries about RDIs



Perliminary tests of .csv files



Outlook

- Enhanced perliminary 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?

Upcomming Questions

- 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?

Future Work

- Improvments 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, .yml etc.)
    - Employ code generating AI
  - Run actual checks on files
    - Return Report with recommendations to user
    - Perform necessary changes to files automatically