Docu-Bot: AI assisted user support

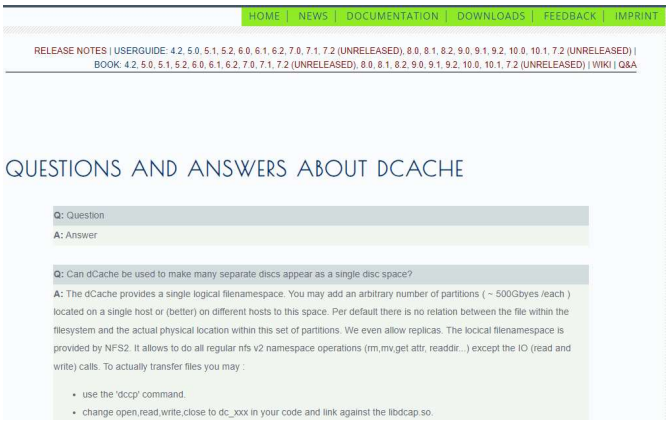
FZU
Institute of Physics of the Czech Academy of Sciences

Jiří Chudoba[1], Michal Chudoba[2], Lukáš Hejtmánek[3]

1 - FZU - Institute of Physics of the Czech Academy of Sciences
2 - Charles University
3 - Masaryk University

Motivation

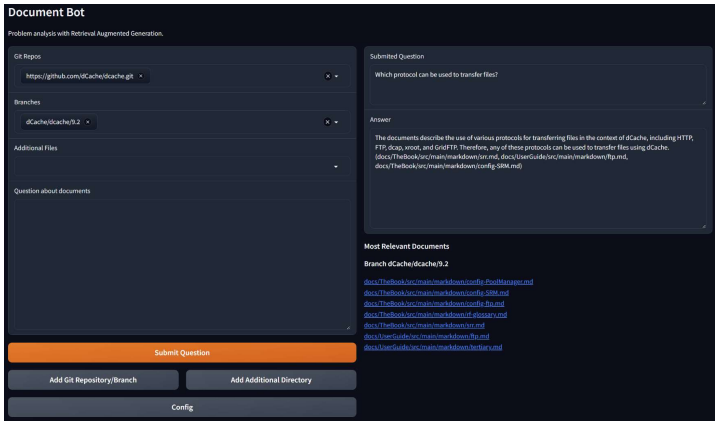
Find answers for questions from users or system administrators about given middleware. Use different online sources of documentation. Check content of available user forums to avoid asking already answered questions.
dCache was our first use case:



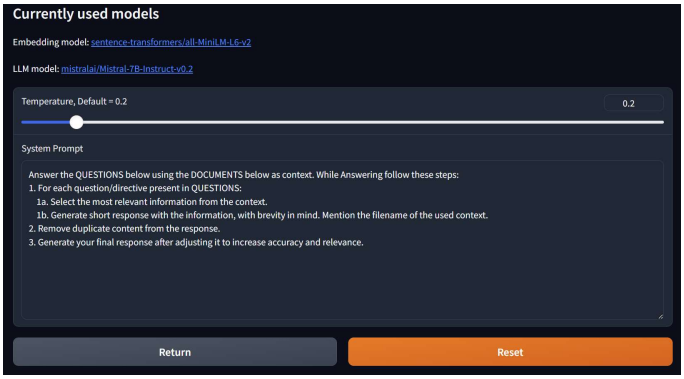
Implementation

- **Utilize online git repositories as a source of information about a product**
- **Add private archive of emails to dcache-forum**
- **Use LLM to generate answers**
  - mistralai/Mistral-7B-Instruct-v0.2
  - version with OpenAI models also available
    - must add payment method to use API, all tests < 10 USD
- **Vectorizer to process questions**
  - sentence-transformers/all-MiniLM-L6-v2
  - OpenAI vectorizer text-embeddings-3-small
- **Custom prompts, more custom documents**

Examples



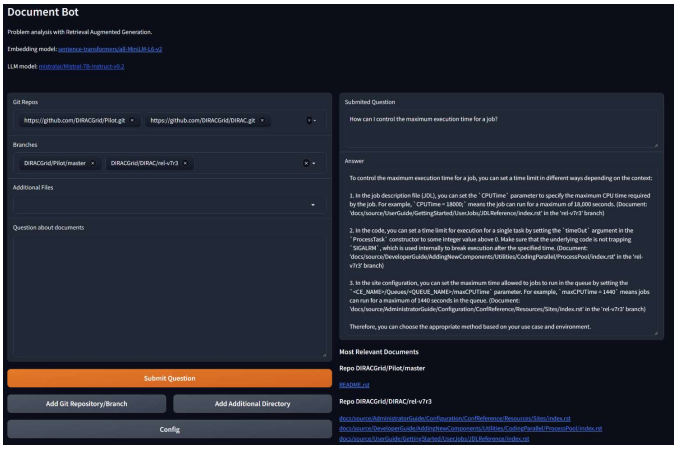
Configuration



Hardware

- **Our default model LLM mistralai/Mistral-7B-Instruct-v0.2 requires at least 8 GB GPURAM**
  - we used NVIDIA T4 with 16 GB GRAM
  - used GRAM: 5.4 GB
  - quantization: LLM 4 bits, processing in 16 bits
- **Bigger model Mixtral-8x7B: 80 GB (NVIDIA A100)**
- **Response time in seconds**
  - 6.5 s (Which protocol can be used to transfer files?)
- **Response when running on CPU only: many tens of seconds**

More examples



Summary

- **Publicly available software for user support improvement**
  - https://github.com/jinymusim/Docu-Bot/
- **Private documentation can be added without sharing it publicly**
- **Dedicated LLM instance quite costly**
  - possibility to share it with many other similar bots
- **Running instance for tests: http://78.128.250.22:7860/**

Computing in High Energy Physics 2024, Cracow

Co-funded by the European Union
MINISTRY OF EDUCATION, YOUTH AND SPORTS
FZU

Institute of Physics of the Czech Academy of Sciences
Na Slovance 1999/2
182 21 Prague 8
Czech Republic
Tel.: +420 266 052 110
Email: secretary@fzu.cz
Web: www.fzu.cz
www.fzu.cz