DocDialog

DocDialog is a *Retrieval Augmented Generation* system for German, focussed on privacy and independence from Cloud providers. It exclusively uses open source AI models, that can be run on local hardware (on premise). DocDialog has an GUI, where the user can upload PDF documents and ask questions to the documents and receive sources containing the answer to the questions asked. The questions can also be answered by the system directly. As generative AI is prone to hallucinations, the answers to the questions may not be correct. The system should be seen as a search system that provides original text snippets that help the user to answer the question.

Installation

Hardware Requirements

The system requires an NVIDIA graphics card with at least 10GB of VRAM.

Software Requirements

The system runs on Linux (Ubuntu). Windows was not tested, but it is possible that the system can run on Windows. You could try to run it in WSL2. If you made it run on Windows, feel free to inform me how to do it, I will put the information here. The following software needs to be installed:

- CUDA (https://docs.nvidia.com/cuda/cuda-installation-guide-linux/ (https://docs.nvidia.com/cuda/cuda-installation-guide-linux/))
- Docker (https://docs.docker.com/engine/install/ubuntu/ (https://docs.docker.com/engine/install/ubuntu/))
- Nvidia-Docker (https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/latest/install-guide.html (https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/latest/install-guide.html))

Key Features

- local hardware only

Quickstart

Clone the git: git clone https://github.com/jtwiefel/doc_dialog (https://github.com/jtwiefel/doc_dialog)

Run the LLM: cd doc_dialog docker/scripts/start_llm.sh

Run the GUI: cd docker/scripts ./start_gui.sh

Open the Browser and go to: http://localhost:7860

Upload some PDFs by navigating to "Dokumente Verwalten" and drop the PDfs into the Dropzone. Click "in Datenbank laden". This will take some time. When the documents are ready, they will be listed in "Hochgeladene Dokumente". You can also delete some of them by clicking "aus Datenbank entfernen". You can but do not need to persist the database by clicking on "Änderungen an Datenbank speichern". This will save your database. When you restart the program, the PDFs will then still be stored in the database. If you want to do only temporary requests, you do not need to persist the changes. Now you can start to ask questions. Navigate back to "Frage stellen". You can now enter a question at "Frage" and click "Frage verarbeiten". The system will search for relevant sources in your documents that help to answer the question and display them under "Quellen". You can also set the checkbox for "Frage beantworten" before clicking "Frage verarbeiten". The system then tries to give an answer to the question below "Antwort". Note that the answer may not be correct and the processing time can be very long (more than 60s).

License