Contribution ID: **29** Type: **Oral Presentation**

RAG chatbot for user and administrator support

Thursday, 20 March 2025 16:00 (20 minutes)

Users may have difficulties finding answers in the documentation for products, when many pages of documentation are available on multiple web pages or in email forums. We have developed and tested an AI based tool, which can help users to find answers to their questions. Our product called Docu-bot uses Retrieval Augmentation Generation solution to generate answers to various questions. It uses github or open gitlab repositories with documentation as a source of information. Zip files with documentation in a plain text or markdown format can also be used for input. Updated version of Docu-bot can also process pdf files.
Embedder model and Large Language Model generate answers. All models conforming to OpenAI python API can be used. For this reason we have developed and deployed a custom LLM and embedder server with Llama-3.2-3B model and all-mpnet-base-v1 model on single Nvidia T4 GPU. This allows the user to actively choose between our model and OpenAI model and allows us to support multiple Docu-bot instances without the need for more compute power. We have experimented and deployed setups with document reranking and chat interface to gauge the correctness and relevancy of responses with the provided filtered context and chat history.

Primary authors: Dr CHUDOBA, Jiri (Institute of Physics of the Czech Academy of Sciences); Mr CHUDOBA, Michal (Charles university, Prague)

Presenter: Dr CHUDOBA, Jiri (Institute of Physics of the Czech Academy of Sciences)

Session Classification: Artificial Intelligence (AI) - II

Track Classification: Track 10: Artificial Intelligence (AI)