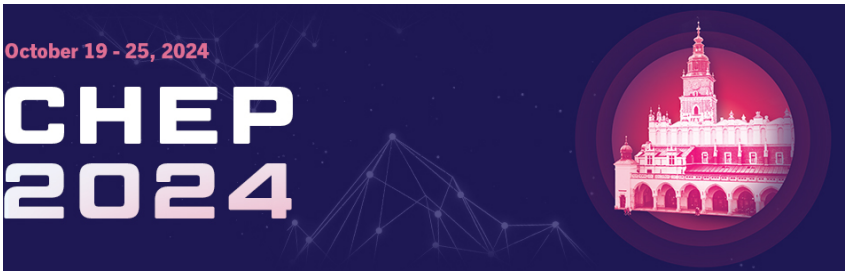

Contribution ID: **493** Type: **Talk**

AccGPT: A CERN Knowledge Retrieval Chatbot

Monday 21 October 2024 17:09 (18 minutes)

In the vast landscape of CERN's internal documentation, finding and accessing relevant detailed information remains a complex and time-consuming task. To address this challenge, the AccGPT project proposes the development of an intelligent chatbot leveraging Natural Language Processing (NLP) technologies. The primary objective is to harness open-source Large Language Models (LLMs) to create a purpose-built chatbot for text knowledge retrieval, with the potential to serve as an assistant for code development and other features in the future.

This initiative was driven by the growing demand at CERN for access to LLMs, not only for building AI Chatbots but also for various other use cases, including Transcription and Translation as a Service (TTaaS), CDS and Zenodo Information Categorization, HR selection processes, and many others. Providing easy and efficient access to LLMs is crucial for the adoption of Generative AI across numerous processes at CERN.

A promising first prototype has already been developed in the realm of knowledge retrieval. It demonstrates a sufficient understanding of user inquiries and provides comprehensive responses utilizing a Retrieval Augmented Generation (RAG) pipeline. However, there is room for improvement to further increase the precision of the responses, which can be achieved by enhancing the retrieval pipeline, considering more powerful and larger LLMs, or fine-tuning the LLMs with more relevant scientific data.

The user interface design and overall user experience of the current prototype chatbot are being iteratively improved, and preparations are underway to make AccGPT available to the community for testing. Automated data scraping and preprocessing pipelines are also being developed to update the chatbot's knowledge base fully autonomously.

Primary authors: Dr REHM, Florian (CERN); GUIJARRO, Juan Manuel (CERN); Dr VALLECORSA, Sofia (CERN); KAIN, Verena (CERN)

Presenter: GUIJARRO, Juan Manuel (CERN)

Session Classification: Parallel (Track 6)

Track Classification: Track 6 - Collaborative software and maintainability