Contribution ID: **15** Type: **not specified**

Leveraging RAG Architecture for Effective Email Response Automation: a CNAF Tier-1 User Support use case

Wednesday, 12 June 2024 10:05 (25 minutes)

CNAF provides computing resources to over 60 scientific communities and supports over 1700 active users through its *User Support* (US) department. US handles daily emails and tickets to help users in employing effectively computing resources and using latest software technologies. Since 2003, CNAF hosts the main INFN computing center, one of WLCG Tier-1.

The primary challenge is to handle user queries and support requests related to the hardware and software technologies employed by the various experiments at CNAF. Users often require expert assistance via email to troubleshoot issues, optimize code performances, or leverage specialized features of the Tier-1 infrastructure.

To tackle this problem, we propose the development of a *Retrieval Augmented Generation* (RAG) model tailored specifically for automating responses to support-related emails. The RAG model will leverage advanced *Natural Language Processing* techniques to understand and generate accurate responses based on the context retrieved and the user queries given.

Through the LangChain framework [1], we have built a vector database containing information from the Tier-1 User Guide [2]. Leveraging an efficient retrieval system [3] and following the RAG pipeline, portions of the User Guide have been prompted directly into a *Large Language Model* (LLM) to address user queries. This approach has aimed to enhance the responsiveness and accuracy of the LLM through specific domain knowledge encoded within vector spaces and cutting-edge semantic similarity models, grasping human-like responses through *Foundation Models*.

References

[1] Applications that can reason. Powered by LangChain, https://www.langchain.com

[2] INFN-CNAF Tier-1 User Guide, https://l.infn.it/t1guide

[3] Wen Li *et al.*, "Approximate Nearest Neighbor Search on High Dimensional Data - Experiments, Analyses, and Improvement."IEEE Trans. Knowl. Data Eng. **32** (2020) 1475

Primary authors: TRASHAJ, Alberto (Università di Bologna); BARBETTI, Matteo (INFN CNAF); RONCHIERI, Elisabetta (Istituto Nazionale di Fisica Nucleare); PELLEGRINO, Carmelo (Istituto Nazionale di Fisica Nucleare); CESINI, Daniele (Istituto Nazionale di Fisica Nucleare); GIUGLIANO, Carmen (Istituto Nazionale di Fisica Nucleare); LATTANZIO, Daniele (Istituto Nazionale di Fisica Nucleare); MORGANTI, Lucia (Istituto Nazionale di Fisica Nucleare); PASCOLINI, Alessandro (Istituto Nazionale di Fisica Nucleare); RENDINA, Andrea (Istituto Nazionale di Fisica Nucleare); SHTIMMERMAN, Aksieniia (Istituto Nazionale di Fisica Nucleare)

Presenter: TRASHAJ, Alberto (Università di Bologna)

Session Classification: Wednesday morning: Part I