PROJECT ABSTRACT

Group 5: Dario Filatrella, Michele Palma, Mattia Scardecchia, Federico Zarantonello

For our project we are going to implement, finetune and evaluate a Retrieval Augmented Generation system that can discuss a range of topics in the STEM domain. To showcase our system, we plan to develop a simple application with a chat interface to explore scientific papers available in the Arxiv database.

We will download, clean and filter the English Wikipedia (~100gb) and build a vector database of semantic embeddings based on STEM articles. We will then build a RAG system using an open source llm such as Gemma 2B. To evaluate our system we will consider its accuracy on question answering benchmarks focusing on the STEM domain, e.g. using an appropriate subset of MMLU.

To improve the performance of our system we plan to finetune both the llm and the retriever on a STEM question answering task, using the recent decoupled RA-DIT approach. We will assess the performance with and without RAG, as well as before and after finetuning, and compare against simple baselines such as BERT (NSP).