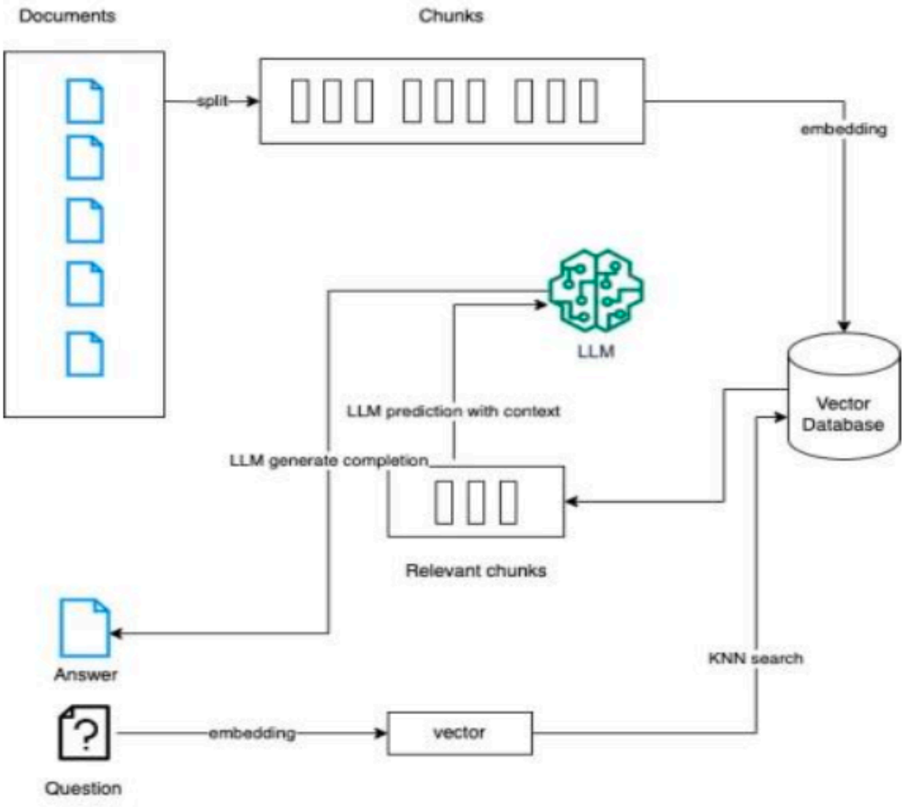
RAG

RAG (Retrieval-Augmented Generation) is an approach of combining the classic *information retrieval* techniques with modern *generative technologies*. It is introduced by Meta as a solution for improving the accuracy of LLM applications by enabling them with access to new and dynamically extendable information, that exists in external, for them, sources.

In RAG, the developer can collect a set of documents, relevant to the domain, to pre-process them independently from the LLM processing, and at the next step, to integrate the RAG pre-processing outcome with the input to the LLM application.

The process requires applying same models of chunking, vectorisation, and embedding of the content of the external documents, as the models used for vectorisation of the prompts - the human questions in question-answering, chat, and other text generation systems.



RAG Architecture, image source

The advantages of involving RAG in the process of NLU are the enabling of personalisation and better adaptation to the tasks, keeping consistency of the LLM by providing it with the new available facts, and therefore ensuring higher reliability of the operations results.

Data Science Holodeck projects applies RAG by aggregating data in a knowledge graph and pre-processing it with graph-based algorithms, before integrating it with LLM for generative language-specific processing.