Retrieval Augmented Generation

Duarte Santos
DETI
Universidade de AVeiro
124376 - duartevsantos@ua.pt

***Abstract*—Retrieval Augmented Generation (RAG) is a model that combines the strengths of retrieval-based and generation-based models (GPT [?], Llama [?]). It uses a retriever to find relevant information from a large corpus and a generator to produce the final output. This model has been shown to outperform previous models in several tasks, such as question answering and text summarization. In this monograph, I present an overview of the RAG model, its architecture, and its training process. It's also discussed the advantages and limitations of the model and present some of the most recent research in the field. Finally, I present some ideas for future research and discuss the potential impact of RAG on the field of natural language processing.**

***Index Terms*—Retrieval Augmented Generation, RAG, Natural Language Processing, Question Answering, Text Summarization**