Project Proposal
Advanced Topics on Machine Learning 2025/2026
MSc Data Science and Engineering (FEUP)

Title: Document Information as Non-Parametric Memory using Dense Passage Retrieval and Custom Reranker Integrated to a LLM.

Team: Danillo Silva (up202300683), Helena Alves (up202403103), Paula Ito (up202308611).

Project idea: The goal is to implement and evaluate a Retrieval-Augmented Generation (RAG) system for open-domain Question Answering. We will use a dataset containing query, positive passage and negative passage examples to fine-tune the dual encoder component. We will compare three different approaches:

#	Approach	Description
1	Baseline (LLM-only)	Directly prompt the LLM to answer the question without external context.
2	Upperline (LLM + golden context)	Directly prompt the LLM to answer the question with the positive (golden) passage.
3	LLM + RAG with fine-tuned DPR Dual Encoder	The **Dual Encoder** will be fine-tuned using our data on a pre-trained encoder model. The **retrieval** of top-K documents will be done using similarity search.
4	LLM + RAG with fine-tuned DPR Dual Encoder + Custom Reranker	Same **Dual Encoder** from approach 2. The **retrieval** of K candidates will be done using similarity search, which will then be refined with a **custom reranker neural network** that will re-score each query-passage pair.

The retrieval performance will be evaluated using standard Information Retrieval metrics, such as Mean Reciprocal Rank @k (k to be defined). The final evaluation on the LLM will be evaluated through question and answering, whose metric is still under refinement, but it will likely be LLM-as-a-Judge.

Datasets: MS MARCO is a large scale information retrieval corpus that was created based on real user search queries using the Bing search engine. It provides millions of Query, Positive Passage and Negative Passage examples, which are the essential labels used for fine-tuning the DPR Dual Encoder model via contrastive learning. It also will be used to retrieve the gold answers for final LLM evaluation.

Papers to read

1. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (RAG) (Lewis et al., NeurIPS 2020). This paper introduced and defined the RAG architecture, demonstrating how to combine a retriever with a seq2seq generator (LLM).
2. Dense Passage Retrieval for Open-Domain Question Answering (DPR) (Karpukhin et al., EMNLP 2020). This is the foundational paper that introduced the Dense Passage Retriever architecture, which serves as the retrieval component in our project.