Industry Project Proposal

Information Retrieval with LLMs

Shubhangi Sanyal, Anurag Dey

Mentors: Sarvesh Bhandaokar, Srijan Bhushan

Introduction

Information Retrieval systems operate on extensive repositories. Hence, the efficiency of retrieval algorithms becomes of paramount importance. To improve the user experience, the retrieval performance is enhanced from both the upstream (query reformulation) and downstream (reranking and reading) perspectives.

The evolution of information retrieval (IR) has progressed from term-based methods and Boolean logic to the integration of neural models. Initially focused on keyword matching, IR transitioned to vector space models, allowing for the capture of semantic relationships. Further advancements included statistical language models, refining relevance estimation through contextual and probabilistic factors. Recently, large language models have been recognized as powerful tools exhibiting remarkable proficiency in language understanding and generation.

Problem Statement

This project aims to explore and implement Language Model-based approaches, specifically Leveraging Pre-trained Language Models (LLMs), to enhance information retrieval systems. Leveraging the power of advanced language models, such as GPT (Generative Pre-trained Transformer) and its variants, this project seeks to transform the traditional information retrieval process by optimizing the search, relevance, and contextual understanding of retrieved information.

Learning Outcomes

Keeping in mind the aim of the project, here are some of the learning outcomes that can be expected from the completion of this project.

- The objective of the Project is to understand and implement LLMs for Information Retrieval.
- Especially, harnessing the advantage of LLMs in each stage of an Information retrieval engine, mainly query rewriting, retrieving, re-ranking, and reading.
- The expected outcome is to create an end-to-end solution that can take internal data as input and create a Personalized Chat Bot while retrieving relevant information from the data.
- This can be done by leveraging semantic understanding, to enhance search capabilities of traditional search engines.
- Further, we will also evaluate & benchmark our prototype against traditional NLP-based approaches to measure the performance gain.
- Certain use cases of this prototype would be resume-shortlisting, chat bots for institution/company, FAQ, etc.

Methodology

- Explore and discuss the Research paper - Large Language Models for Information Retrieval: A Survey.
- Create a basic Information Retrieval Solution trained on a corpus of CV data, and evaluate its performance.
- Explore various state-of-the-art LLMs (LLaMA-2, FalconLLM, GPT, etc) and retrain those on the same corpus of data.
- Evaluate model performance and gain against the traditional NLP approach.
- Create an end-to-end pipeline solution: Data processing, model training/transfer learning, hyperparameter tuning, Output presentation (basic UI).
- Replicate & deploy the pipeline solution on different data corpus and finetune the pipeline to create a stable generic solution.

References

- Research paper - Large Language Models for Information Retrieval: A Survey (2308.07107, arxiv.org)
- Example of a simple Information Retrieval Engine run on Python code (Google Colab)
- Research paper on query expansion: Query2doc (2303.07678, arxiv.org)
- Examples of re-ranking in IR using a large language model (LLMReranker-Lyft-10k, LLMReranker-Gatsby)