# Internship Report

Aditya Ghosh

Under The Supervision of

Dr Debapriyo Majumdar
Computer Vision and Pattern Recognition Unit
At Indian Statistical Institute, Kolkata

From: 6th June to 10th July, 2024

## Models used:

The primary model utilized was the Mistral-7B version 0.1 (mistralai/Mistral-7B-v0.1), which was quantized to a 4-bit precision version to achieve faster inference times.

For summarization tasks, the facebook/bart-large-cnn model was employed.

Additionally, Sentence Transformers (commonly referred to as SBERT) were used to create sentence embeddings for assessing similarity between sentence pairs.

Finetuning of the Mistral-7B model was accomplished using Low-Rank Adaptation (LoRA) adapters. The model was optimized for efficient k-bit training by configuring LoRA for specific layers, and the training environment was set up using Hugging Face's Trainer module.

## Query expansion into description question:

The objective was to expand a short query, typically consisting of 2-4 words, into a more detailed question that the query might be asking.

The datasets trec-web-2009, robust-2004 and trec-web-2002 (clueweb09/catb/trec-web-2009, disks45/nocr/trec-robust-2004, gov/trec-web-2002) were used for training the finetuned mistral-7b model to generate a longer description question from a shorter query.

The prompt used to finetune the model was:

"*Given a small query give me what the bigger questions the smaller query could have come from.*

*### Query:*

*{data_point["query"]}*

*### Description:*

"

10% of the dataset was set aside as a test set for evaluation purposes. Sentence embeddings were generated for both the test set and the model-generated sentences using Sentence Transformers (SBERT). The similarity between these sentence embeddings was then calculated. The results varied between 0.930699 and 0.391612, with an average similarity score of 0.721828. The first five results are as follows:

| query | mistral_description | dataset_description | score |
|---|---|---|---|
| rick warren | Find information about Rick Warren, the author of the book "The Purpose Driven Life." | I'm looking for information on Rick Warren, the evangelical minister. | 0.5179799795150757 |
| British Chunnel impact | What has been the impact of the British Chunnel on the economy of the UK and France? | What has been the","What impact has the Chunnel had on the British economy and/or the life style of the British?" | 0.8396186232566833 |
| oceanographic vessels | Find documents that describe oceanographic vessels. | Identify documents that discuss the activities or equipment of oceanographic vessels. | 0.9267082214355469 |
| used car parts | Find information on where to buy used car parts. | I am looking for sources for parts for cars, preferably used. | 0.7555357217788696 |
| orphan drugs | Find documents that discuss the development of drugs for orphan diseases. | Find documents that discuss issues associated with so-called "orphan drugs", that is, drugs that treat diseases affecting relatively few people. | 0.8916334509849548 |

# What did not work:

## Finetuning mistral-7b to predict the term frequencies and inverse document frequencies of the query:

The TREC-Robust-04 dataset (disks45/nocr/trec-robust-2004) was used for this task. An initial issue was that the model consistently predicted term frequencies as whole numbers (1.0 or 2.0), likely because these values caused the least error during evaluation. Consequently, term frequencies and inverse document frequencies (TF-IDF) were not properly calculated based on the actual document-query terms and query length.

Initially, TF-IDF values were manually calculated using standard formulas. However, it was later discovered that these values could be directly queried from Pyserini, which provided a faster and more efficient solution. As a result, term frequency and inverse document frequency retrieval was shifted to Pyserini.

## Reranking documents based on how relevant they are to the query:

Another approach that was attempted involved using the Pyserini Python package to retrieve highly relevant documents from the robust-04 dataset based on a given query, followed by an attempt to re-rank the documents by relevancy.

Initially, the plan was to feed document-query pairs into the Mistral-7B model and request a relevancy score between 0 and 10, based on how closely the document matched the query. However, this method proved ineffective when applied to the top retrieved documents using BM-25 or other traditional information retrieval techniques. The model consistently assigned high scores of 9 and 10 to all documents, as they were generally relevant to the query. As a result, it failed to distinguish between these highly ranked documents and determine which was more relevant. Consequently, this approach was abandoned.

## Summarising document pairs to check which document is more relevant:

In the next phase, a process was developed for summarizing document pairs to assess which was more relevant to a given query. Summaries were generated for each document and then fed into the Mistral-7B model, which was used to evaluate the relevance of the summaries. To accommodate the model's input limit, documents were split into smaller segments of 4096 tokens.

The facebook/bart-large-cnn model, a transformer-based model fine-tuned on CNN news articles by Facebook, was selected for summarization. It was chosen based on its high performance and popularity on Hugging Face, standing out among the top-rated summarization models.

The primary challenge that was encountered was the extended inference time required for each document chunk, leading to inefficiencies when multiple documents had to be summarized and re-ranked. Ultimately, this led to this approach being abandoned entirely.