Executive Summary

Summary of Methodologies

This project builds a specific pipeline of Retrieval-Augmented Generation (RAG) for a question answering machine using Reddit comments using the following techniques**:**

- **Synthetic query generator**
- **SBERT embedding (vectorization of sentences)**
- **K-Means Clustering**

We also develop the following evaluation metrics that is suitable to test the quality of the retrieval:

- **Cosine Precision**
- **Ranked Cosine Precision**

Summary of Results

- Our clustering method reduces the runtime of retrieval significantly
- We vectorize the (Reddit) comments and the given query using SBERT and rank the comments by cosine similarities to the query (i.e., naive RAG)
- We generate similar queries to the original query using the synthetic query generator and re-rank the sentences by averaging the cosine similarities to all the queries
- Our averaging method yields a better retrieval with either method