Evaluation Report

Evaluation 1: Comparing the effectiveness of different chunking strategies and embedding models based on retrieval time and accuracy.

Findings:

Embdedding Method	**Chunking Method**	**Retrieval Time**
all-MiniLM-L6-v2	RecursiveCharacterTextSplitter	1.80s
all-MiniLM-L6-v2	CharacterTextSplitter	1.66s
multi-qa-MiniLM-L6-cos-v1	RecursiveCharacterTextSplitter	0.36s
multi-qa-MiniLM-L6-cos-v1	CharacterTextSplitter	1.46s

Observation:

- **Best Embedding Model**: *multi-qa-MiniLM-L6-cos-v1* outperformed *all-MiniLM-L6-v2* in retrieval speed across both chunking methods.
- **Best Chunking Method:** *RecursiveCharacterTextSplitter* was fastest when paired with *multi-qa-MiniLM-L6-cos-v1* **(0.36s)**.
- **Overall Best Setup:** *multi-qa-MiniLM-L6-cos-v1 + RecursiveCharacterTextSplitter* yielded the best retrieval time **(0.36s)**.

Evaluation 2: Comparing the two different similarity search algorithms (i.e. cosine, euclidean) based on retrieval time.

Findings:

The Evaluation 1 is done under cosine similarity search. The following is done under Euclidean distance similarity search.

For the calculation of retrieval accuracy, I chose the **MRR (Mean Reciprocal Rank)** metric which is used to evaluate the systems that return a list of ranked results such as document retrievers.

Formula:

$$\mathrm{MRR} = \frac{1}{\mathrm{Q}} \sum_{i=1}^{\mathrm{Q}} \frac{1}{\mathrm{rank}_{\mathrm{i}}}$$

Where,

Q = No. of queries

$\mathrm{rank}_{\mathrm{i}}$ = Position of 1st related document in list for the query i

Calculation:

1st combination (embedding = multi-qa-MiniLM-L6-cos-v1, chunking = RecursiveCharacterTextSplitter)

Query	Rank
Q1	1
Q2	1
Q3	1
Q4	2
Q5	1

Applying the MRR formula:

$$MRR = \frac{1}{5} \sum_{i=1}^{5} \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{2} + \frac{1}{1}$$

We get, MRR = 0.9

Similarly, doing the same for other 3 combination, we get,

Embdedding Method	**Chunking Method**	**Retrieval Accuracy**
all-MiniLM-L6-v2	RecursiveCharacterTextSplitter	90%

all-MiniLM-L6-v2	CharacterTextSplitter	80%
multi-qa-MiniLM-L6-cos-v1	RecursiveCharacterTextSplitter	90%
multi-qa-MiniLM-L6-cos-v1	CharacterTextSplitter	90%

Observation:

I tested 5 queries on each combination of embedding method (all-MiniLM-L6-v2 and multi-qa-MiniLM-L6-cos-v1) and chunking method (RecursiveCharacterTextSplitter and CharacterTextSplitter) and Euclidean distance for similarity search.

The results show that:

- **multi-qa-MiniLM-L6-cos-v1** consistently achieved **high retrieval accuracy (90%)** across both chunking methods. This suggests that the model is robust to the choice of chunking strategy when using Euclidean distance.
- **all-MiniLM-L6-v2** showed more variability:
    - It achieved only **80% accuracy** when paired with RecursiveCharacterTextSplitter, indicating that the chunking method may influence performance more significantly for this embedding model under Euclidean distance.
    - However, with CharacterTextSplitter, it reached **90% accuracy**, matching the performance of multi-qa-MiniLM-L6-cos-v1.