There are 3 approaches that we adopted as part of data cleaning/preprocessing:

1. Using raw data to evaluate the performance of retriever
2. Adding preprocessed tables with meta data
3. Using c4 data cleaning strategies

For each of these approaches, we carry out the following steps:

* Chunking
* Parrot paraphrasing
* Ensemble retrieval
* cohereRerank

Note: In cohereRerank, we set a threshold value of 0.85 for the relevance score. If the score was greater than this value then the document was deemed relevant, and vice versa. It was also observed that this score was misleading in some iterations therefore, manual checking for the presence of relevant documents also contributed to finalizing the results below.

Approach 1:

|  |  |  |
| --- | --- | --- |
| Chunk\_size, overlap | Character text splitter | Recursive text splitter  Default separators  ["\n\n", "\n", " ", ""] |
| 1000, 200 | 5/6 with ‘.’  6/6 with ‘\n’ | 6/6 |
| 500, 100 | 6/6 with ‘\n’  6/6 with ‘.’ | 6/6 |
| 700, 200 | 6/6 with ‘.’  6/6 with ‘\n’ | 6/6 |

Semantic chunking:

4/6 for percentile

4/6 for std

5/6 for iqr

Approach 2:

|  |  |  |
| --- | --- | --- |
| Chunk\_size, overlap | Character text splitter | Recursive text splitter  Default separators  ["\n\n", "\n", " ", ""] |
| 1000, 200 | 5/6 with ‘.’  6/6 with ‘\n’ | 6/6 |
| 500, 100 | 6/6 with “.”  6/6 with “\n” | 6/6 |
| 700, 200 | 6/6 with “.”  6/6 with “\n” | 6/6 |

Semantic Chunking:

4/6 with percentile

5/6 with std

4/6 with iqr

Approach 3:

|  |  |  |
| --- | --- | --- |
| Chunk\_size, overlap | Character text splitter | Recursive text splitter  Default separators  ["\n\n", "\n", " ", ""] |
| 1000, 200 | 5/6 with ‘.’  6/6 with ‘\n’ | 6/6 |
| 500, 100 | 6/6 with ‘.’  6/6 with ‘\n’ | 6/6 |
| 700, 200 | 6/6 with ‘.’  6/6 with ‘\n’ | 6/6 |

Semantic chunking:

6/6 with percentile

Std and iqr took over 30 minutes to run

Analysis:

Semantic chunking in approach 3 might have produced better results than in any other approach, but since it was taking too long to create chunks, we had to halt the process.