Evaluation

To understand the effectiveness and behaviour of the BM25 and all-MiniLM-L6-v2 retrieval pipeline, a three-part evaluation was designed. 150 queries were generated using GPT-4 (reference) to test on the model and produce statistics on its performance. These queries were split into the following three categories:

1.) Matching queries - 50 queries based on content that appears in the dataset, each paired with a “ground truth” summary. The ground truths were the summaries that best answered the queries according to GPT-4 (2023 [1]). For example: “What is The North Island brown kiwi?”

2.) Non-matching queries - 50 queries that had nothing to do with the dataset. These did not feature ground truth, as they were designed to lack a satisfying answer in the partition of the Qulac dataset (Qualc [2]) used in the pipeline. For example, “Who won the first season of The Voice?”

3.) Ambiguous queries - 50 queries that were generalisations of topics related to summaries in the dataset but not directly answerable easily. For example, “Tell me about famous political figures.”

All queries were run through the full BM25 (Brown et. Al [3]) and all-MiniLM-L6-v2 (Hugging Face [4]) pipeline. In the case of the 50 matching queries, the summary chosen by the pipeline was compared against the ground-truth using semantic similarity with util.cos\_sim (Sentence Transformers [5]). If the similarity between the model’s summary and the ground truth summary was greater than or equal to 0.7, the retrieval was considered successful. This allowed for instances where there were two or more suitable summaries available in the dataset for a given query. Additionally, classification metrics were computed. Confident predictions were treated as positive predictions, and semantic matches defined correctness.

References:  
[1] <https://openai.com/index/gpt-4-research/>

[2] Whatever reference is being used for the Qulac dataset

[3] <https://zenodo.org/records/6106156>

[4] <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

[5] https://sbert.net/docs/package\_reference/util.html