Retrieval Pipeline Results

Matching queries results:

Accuracy: 92.00%

Precision: 100.00%

Recall: 91.49%

F1 Score: 95.56%

Avg. cosine (query vs prediction): 0.5882

Avg. semantic similarity (prediction vs expected): 0.9594

These results indicate strong metrics across the board, particularly in precision, meaning the system was never confidently wrong. However, it occasionally failed to respond confidently even when it should have as seen by the slightly lower recall. This suggests the pipeline was cautious yet reliable.

To comment on the logic used to compute these statistics, the following definitions were coded into the evaluation script:

|  |  |
| --- | --- |
| Behaviour | Term |
| Confident and semantic similarity ≥ 0.7 | True Positive (TP) |
| Confident but semantic similarity < 0.7 | False Positive (FP) |
| Not confident but semantic similarity ≥ 0.7 | False Negative (FN) |
| Not confident and semantic similarity < 0.7 | True Negative (TN) |

To calculate the evaluation metrics, the following code was used:

A screenshot of a computer program

AI-generated content may be incorrect.

It should be noted that True Negatives were set to 0 as every query had a correct associated answer, meaning that there could be no True Negatives for the matching queries.

As for the ambiguous queries, the following results were produced:

Confident predictions: 42%

Avg. cosine similarity: 0.3584

In this case, the model was appropriately hesitant. The low average demonstrates that even when the pipeline attempted an answer (which happened 42% of the time), the fit was often poor (0.3584 average cosine similarity), which supports the system’s ability to avoid overconfident guessing where the model is not sure.

As for the non-matching queries, the metrics seen below were calculated:

Correctly cautious (not confident): 80%

False positives (wrongly confident): 20%

Avg. cosine similarity: 0.2839

This indicates the system could not answer 80% of irrelevant queries, which is desirable when compared to matching and ambiguous queries. However, the 20% false positive rate means the system attempted to answer one in five queries, even when no relevant summaries existed. This is disappointing, but relatively fine compared to the confidence scores from the other query groups (92% and 42% respectively).

Ultimately, the evaluation of the retrieval pipeline was robust, and highlighted the success of the model. However, it was not possible to implement Precision, Recall, and F-1 scores for the ambiguous and non-matching query groups, as ground truths were impossible to even synthetically produce due to the nature of the groups. Ideally, comparing the model’s results against the results of a perfect model would be the best evaluation method, but the task is to subjective to expect perfection from any pipeline. The drops in accuracy and cosine similarity as the model transitioned from matching, to ambiguous, to non-matching do represent a solid evaluation architecture though, and suggest the pipeline performance was more than capable of performing summary retrieval where fit for any given query.