#### **Overview**

This report summarizes the RAG model’s performance based on the average accuracy for the following four key metrics: **Retrieval Accuracy**, **Model Accuracy**, **ROUGE-1**, and **ROUGE-L.**

**Here is a metric for ROUGE score for note.** [**https://klu.ai/glossary/rouge-score**](https://klu.ai/glossary/rouge-score)

| **ROUGE Metric** | **Excellent** | **Good** | **Moderate** |
| --- | --- | --- | --- |
| ROUGE-1 | 0.5+ | >0.5 | 0.4-0.5 |
| ROUGE-2 | - | >0.4 | 0.2-0.4 |
| ROUGE-L | - | ~0.4 | 0.3-0.4 |

**Retrieval Accuracy – 2.4/3**

* Interpretation: The model performs well in retrieving relevant documents for each query, with a 2.4 out of 3 average score. This shows that the retrieval component is robust and reliably surfaces content that aligns with query intent. However, it has to be fine tuned for certain keywords - for example using the word mobile causes it to retrieve random user stories not related to user stories at all. Occasionally slightly off target in user story retrieval bringing up 2 extra stories when only 1 was asked for.

**Model Accuracy – 2.13/3**

* Interpretation: The model could benefit from further fine-tuning to improve alignment with user intent and increase response quality, especially for nuanced or complex queries. Feels very hit or miss with more complicated queries, utilizes retrieved chunks to a full degree in most cases.

**ROUGE-1 Score – 0.313**

* ROUGE-1 measures the overlap of single words between the generated and reference text, essentially assessing how many keywords are captured in the output. A score of 0.313 shows partial coverage, meaning that while the model includes some essential terms, it may miss others.

**ROUGE-L Score – 0.285**

* ROUGE-L evaluates the longest common subsequence, capturing the sequence similarity between generated text and the reference.

**Limitations/Possible upgrades:**

* Allow for memory updates or store the last 3 responses, to allow for follow up questions or to clarify further when the wrong user story is retrieved.
* Inconsistency on whether the summary comes before or after the explanation if there was a way to ensure it always comes first would be better.
* My reference responses might have been not domain specific, application of data augmentation after further understanding.
* Fine tune for domain specific data.

IGS-ht-RAG

* Overall most responses are generated better, However few queries resulted in horrible responses with a 1,1 score they are listed below:
* Is there anything about fitness?
* What are all the mobile related user stories?
* Is there any user story with an assumption that includes internet connection?
* How do i calculate a BLEU Score?

Most likely an issue with the prompt in model.py and query retrieval for the first two.

Otherwise all prompts were replied too perfectly.

As a student how can i stay on top of my goals?

The most improtant features are : {'mind\_maps': ['overview', 'resilience'], 'test\_scenarios': ['Reliability Testing', 'Performance Testing']}

The mongoquery : {'mind\_maps.overview': 1, 'mind\_maps.resilience': 1, 'story': 1, 'refined': 1, 'test\_scenarios.Reliability Testing': 1, 'test\_scenarios.Performance Testing': 1}

Number of retrieved documents from vector DB : 0

Traceback (most recent call last):

File "/mnt/newdisk/experiment/Mongo\_RAG/igs-ht-rag/tester.py", line 44, in <module>

query\_and\_store\_results("input.txt", "output.csv", database\_name, collection\_name, milvus\_collection\_name)

File "/mnt/newdisk/experiment/Mongo\_RAG/igs-ht-rag/tester.py", line 24, in query\_and\_store\_results

mongo\_user\_story\_ids = retrieve\_user\_stories(query, milvus\_collection\_name, 0.2)

File "/mnt/newdisk/experiment/Mongo\_RAG/igs-ht-rag/main.py", line 71, in retrieve\_user\_stories

rank\_scores = reranker.compute\_score(data, normalize=True)

File "/mnt/newdisk/experiment/Mongo\_RAG/env/lib/python3.10/site-packages/FlagEmbedding/abc/inference/AbsReranker.py", line 185, in compute\_score

if isinstance(sentence\_pairs[0], str):

IndexError: list index out of range