The table shows that the system performs as expected when using different embedding models. The expectation being that the system should perform identically when not using RAG and differently with RAG. We see an improvement with the sentence level embeddings (sentence-transformers/sentence-t5-base) when compared to google/flan-T5. I suspect the deterioration in performance from the flan-t5 model to be due to it embedding at a token level. Again, my hypothesis is, this could be due to a higher degree of semantic retention in embedding models that take in larger context.