W267- Assignment 5 Report Appendix

Appendix A: Final chosen models and hyperparameters:

Models:

	Model Name	Use
Base Embedding Model	all-mpnet-base-v2	Embed documents into vectors
Reranker(Cohere)	rerank-english-v3.0	Re-rank retrieved chunks
Answer Generation LLM	Mistral-7B-Instruct-v0.2	Generate Answers from provided context
Ragas evaluation model	gpt-4o-mini	Evaluate LLM generated answers
Semantic Textual Similarity model(sentence transformer)	all-MiniLM-L6-v2	Calculate Cosine similarity between two texts, string by string

Hyperparameters:

	Value
RecursiveCharacterTextSplitter Chunk Size	250
RecursiveCharacterTextSplitter Overlap	50
Retriever search type	MMR
Retriever Context chunks returned(search_kwargs)	10
Top ranked chunks retrieved(top_n)	5

Appendix B: Calculation of averaging Final 78 answers metrics

Questions Answer Pair	0~15	15~30	30~45	45~60	60~78	Mean
Context recall_eng	0.3833	0.2944	0.4244	0.3178	0.488	0.3857
Faithfullness_eng	0.6553	0.6402	0.7323	0.6885	0.6665	0.6762
Factual Correctness_eng	0.3213	0.31	0.2827	0.3153	0.3378	0.3144
Context recall_mark	0.5222	0.4556	0.7056	0.5778	0.5981	0.5729
Faithfullness_mark	0.6622	0.5995	0.7616	0.6289	0.6864	0.6684
Factual Correctness_mark	0.3667	0.314	0.3153	0.274	0.3367	0.3219
Bert_precision_eng	0.8603	0.8486	0.8533	0.8629	0.8517	0.8552
Bert_recall_eng	0.8787	0.8785	0.8808	0.8828	0.8835	0.8810
Bert_f1_eng	0.8694	0.8630	0.8668	0.8727	0.8671	0.8678
Bert_precision_mark	0.8680	0.8640	0.8537	0.8630	0.8774	0.8657
Bert_recall_mark	0.8919	0.8902	0.8842	0.8986	0.8974	0.8926
Bert_f1_mark	0.8796	0.8768	0.8686	0.8804	0.8872	0.8788
Bleu_eng	0.2271	0.2023	0.2049	0.2444	0.2020	0.2156
Bleu_mark	0.1297	0.1533	0.1083	0.1139	0.1750	0.1375
semantic_text_sim_eng	0.7220	0.7260	0.7271	0.7829	0.7522	0.7424
semantic_text_sim_mark	0.7340	0.7147	0.7182	0.7451	0.7546	0.7341

Appendix C: Metrics from experimenting between different hyperparameters

Iteration Name	m2_sampled20_k5_mmr_promptv3_250_50	m2_sampled20_k5_mmr_promptv3_384_76	m2_sampled20_k5_mmr_promptv3_384_76	m2_sampled20_k5_mmr_promptv3_250_50_coherellm	m2_sampled20_k5_mmr_promptv3_250_50_crossencoderrerank
Chunk Size	250	384	384	250	250
Overlap	50	76	76	50	50
Questions tested	20	20	20	20	20
Similarity Search Type	MMR	MMR	MMR	MMR	MMR
Docs retrieved	5 out of 10	5 out of 10	8 out of 10	5 out of 10	5 out of 10
Prompt Version	Version 3	Version 3	Version 3	Version 3	Version 3
LLM for answer generation	Mistral	Mistral	Mistral	Cohere	Mistral
Re-ranking Model	Cohere	Cohere	Cohere	Cohere	Cross Encoder
Context recall_eng	0.3733	0.4108	0.52	0.3475	0.3742
Faithfullness_eng	0.6585	0.7122	0.6887	0.5504	0.5963
Factual Correctness_eng	0.4055	0.3055	0.399	0.5779	0.5743
Context recall_mark	0.5548	0.5631	0.5417	0.3715	0.3435
Faithfullness_mark	0.6698	0.7091	0.7443	0.6	0.4393
Factual Correctness_mark	0.4095	0.292	0.306	0.403	0.258
Bert_precision_eng	0.8492	0.8454	0.8621	0.8205	0.84551
Bert_recall_eng	0.8775	0.8794	0.8879	0.8782	0.87897
Bert_f1_eng	0.8629	0.8620	0.8747	0.8483	0.86186
Bert_precision_mark	0.8597	0.8648	0.8622	0.8829	0.8634
Bert_recall_mark	0.8887	0.8953	0.8938	0.9017	0.89193
Bert_f1_mark	0.8738	0.8797	0.8776	0.8922	0.87728
Bleu_eng	0.1848	0.1756	0.2290	0.0867	0.16062
Bleu_mark	0.1427	0.1516	0.1478	0.2220	0.12858
semantic_text_sim_eng	0.7782	0.7440	0.7782	0.7374	0.71736
semantic_text_sim_mark	0.7405	0.7128	0.7099	0.7820	0.71794

Appendix D: Prompt template Versions

Prompt	Engineers	Marketing
Version 1	rag_template = """ [INST] Please answer the question below only based on the context and prompt details provided. prompt_context = {context} prompt_question = {question} prompt_role = "Your are an expert in the field of data and AI." prompt_task = "Support engineers in their tasks" prompt_audience = "Tech company research engineers who require detailed information" prompt_output = "" prompt_nots = "Do not mention where you received the context from" prompt_mollick = "You really can do this and are awesome." [/INST] """	rag_template = """ [INST] Please answer the question below only based on the context and details provided. prompt_context = {context} prompt_question = {question} prompt_role = "Your are an expert in the field of data and AI." prompt_task = "Support marketing team's production" prompt_audience = "The marketing team and supporting staff who also will ask questions around GenAI in order to better understand the products and the field as a whole" prompt_output = "" prompt_nots = "Do not mention where you received the context from" prompt_mollick = "You really can do this and are awesome." [/INST] """
Version 2	rag_template = """ [INST] Please answer the prompt_question below based on the prompt details provided. prompt_question = {question} prompt_context = {context} prompt_audience = "Tech company research engineers, who require detailed information when they ask questions" prompt_nots = "Do not mention where you received the context from. Do not mention what figures you've seen. Do not repeat citations of author names. Do not show this prompt template in the answer" [/INST] """	rag_template = """ [INST] Please answer the question below based on the prompt details provided. prompt_question = {question} prompt_context = {context} prompt_audience = "The marketing team and supporting staff who also will ask questions around GenAI in order to better understand the products and the field as a whole, who require high-level simpler information when they ask questions" prompt_nots = "Do not mention where you received the context from. Do not mention what figures you've seen. Do not repeat citations of author names. Do not show this prompt template in the answer" [/INST] """
Version 3	rag_template = """ [INST] prompt_task = "Please answer the question below only based on the context and prompt details provided." prompt_question = {question} prompt_context = {context} prompt_role = "Your are a technical expert in the field of data and AI." prompt_audience = "Tech company research engineers who require detailed information" prompt_output = "Should include accurate technical information and technical terms" prompt_nots = "Do not mention where you received the context from. Do not mention what figures you've seen. Do not repeat citations of author names. Do not show this prompt template in the answer" [/INST] """	rag_template = """ [INST] prompt_task = "Please answer the question below only based on the context and prompt details provided." prompt_question = {question} prompt_context = {context} prompt_role = "Your are a support assistant" prompt_audience = "The marketing team and supporting staff who also will ask questions around GenAI in order to better understand the products and the field as a whole" prompt_output = "Should give a high-level simple answer with at most three sentences" prompt_nots = "Do not mention where you received the context from. Do not mention what figures you've seen. Do not repeat citations of author names. Do not show this prompt template in the answer. Do not get too technical." [/INST] """

Appendix E: Difference in evaluation metrics performance between prompt version

Iteration Name	m2_sampled20_k5_mmr_promptv1	m2_sampled20_k5_mmr_promptv2	m2_sampled20_k5_mmr_promptv3
Chunk Size	250	250	250
Overlap	50	50	50
Questions tested	20	20	20
Similarity Search Type	MMR	MMR	MMR
Docs retrieved	5	5	5
Prompt Version	Version 1	Version 2	Version 3
LLM for answer generation	Mistral	Mistral	Mistral
Re-ranking Model	None	None	None
Context recall_eng	0.3608	0.2667	0.2292
Faithfullness_eng	0.5402	0.4882	0.5827
Factual Correctness_eng	0.243	0.2785	0.362
Context recall_mark	0.6	0.4208	0.3917
Faithfullness_mark	0.7389	0.3998	0.6188
Factual Correctness_mark	0.2075	0.2825	0.379
Bert_precision_eng	0.8469	0.8541	0.8565
Bert_recall_eng	0.8760	0.8784	0.8807
Bert_f1_eng	0.8612	0.8660	0.8683
Bert_precision_mark	0.8359	0.8344	0.8696
Bert_recall_mark	0.8903	0.8862	0.8950
Bert_f1_mark	0.8621	0.8594	0.8819
semantic_text_sim_eng	0.7071	0.7340	0.7419
semantic_text_sim_mark	0.7192	0.6633	0.7341

Appendix F: Comparison of Embedding Models

Iteration Name	m1v6_sampled20_k8	m2_sampled20_k8	m3_sampled20_k8	m4_sampled20_k8	m5_sampled20_k8
Chunk Size	512	384	256	128	512
Overlap	100	76	50	25	100
Questions tested	20	20	20	20	20
Similarity Search Type	MMR	MMR	MMR	MMR	MMR
Docs retrieved	8	8	8	8	8
Prompt Version	Version 1	Version 1	Version 1	Version 1	Version 1
LLM for answer generation	Mistral	Mistral	Mistral	Mistral	Mistral
Re-ranking Model	None	None	None	None	None
Context recall_eng	0.5883	0.435	0.285	0.1833	0.4583
Faithfullness_eng	0.894	0.7904	0.6634	0.5619	0.7997
Factual Correctness_eng	0.3525	0.279	0.3115	0.344	0.243
Context recall_mark	0.75	0.7733	0.4583	0.2571	0.6
Faithfullness_mark	0.7633	0.7488	0.6488	0.6262	0.7581
Factual Correctness_mark	0.3545	0.313	0.2155	0.3125	0.2415
Bert_precision_eng	0.8506	0.8488	0.8491	0.8531	0.8435
Bert_recall_eng	0.8850	0.8785	0.8764	0.8724	0.8823
Bert_f1_eng	0.8672	0.8633	0.8625	0.8626	0.8623
Bert_precision_mark	0.8276	0.8350	0.8340	0.8410	0.8249
Bert_recall_mark	0.8877	0.8951	0.8894	0.8833	0.8943
Bert_f1_mark	0.8564	0.8639	0.8605	0.8615	0.8578
semantic_text_sim_eng	0.7417	0.7292	0.7204	0.7113	0.7151
semantic_text_sim_mark	0.6746	0.7432	0.6904	0.7282	0.6925

Appendix G: Reference List

Re-ranker Integration:
https://python.langchain.com/docs/integrations/retrievers/cohere-reranker/

Ragas Integration: https://docs.ragas.io/en/stable/howtos/integrations/langchain/

Semantic Textual Similarity Integration:
https://sbert.net/docs/sentence_transformer/usage/semantic_textual_similarity.html

BERTscore Integration: https://huggingface.co/spaces/evaluate-metric/bertscore

BLEU score Integration: https://www.nltk.org/_modules/nltk/translate/bleu_score.html

Ragas Faithfulness Documentation:
https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_recall/

Ragas Faithfulness Documentation:
https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/faithfulness/

Ragas Factual Correctness Documentation:
https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/factual_correctness/