Evaluation Results Summary

Optimization

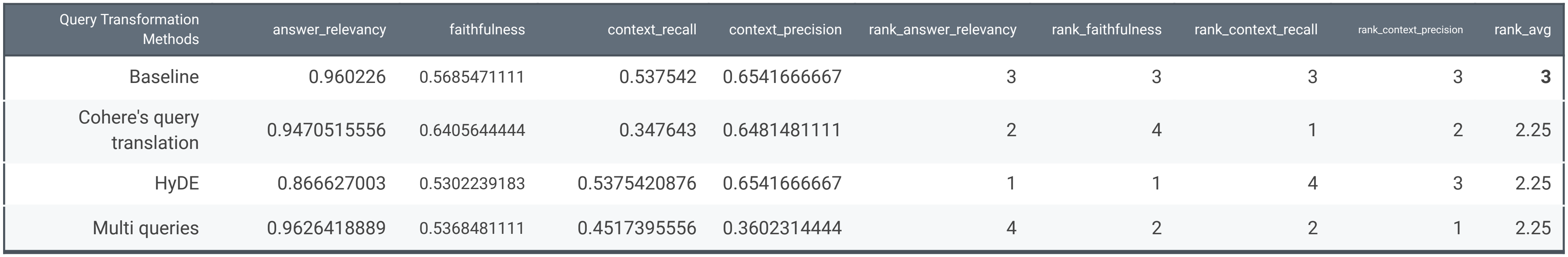
1. Prompt Template

Prompts ID	answer_relevancy	faithfulness	context_recall	context_precision	answer_relevancy_rank	faithfulness_rank	context_recall_rank	context_precision_rank	rank_avg
Base Template	0.701	**0.6667**	0.5714	0.6071	2	5	1	1	2.25
Template 2	0.5614	0.5391	0.5714	0.6627	1	2	1	2	1.5
Template 3	**0.9765**	0.5938	0.5714	0.6905	5	3	1	4	**3.25**
Template 4	0.9735	0.4973	0.5714	**0.7024**	4	1	1	5	2.75
Template 5	0.9715	0.61	0.5714	0.6627	3	4	1	2	2.5

2. Chunking Strategies

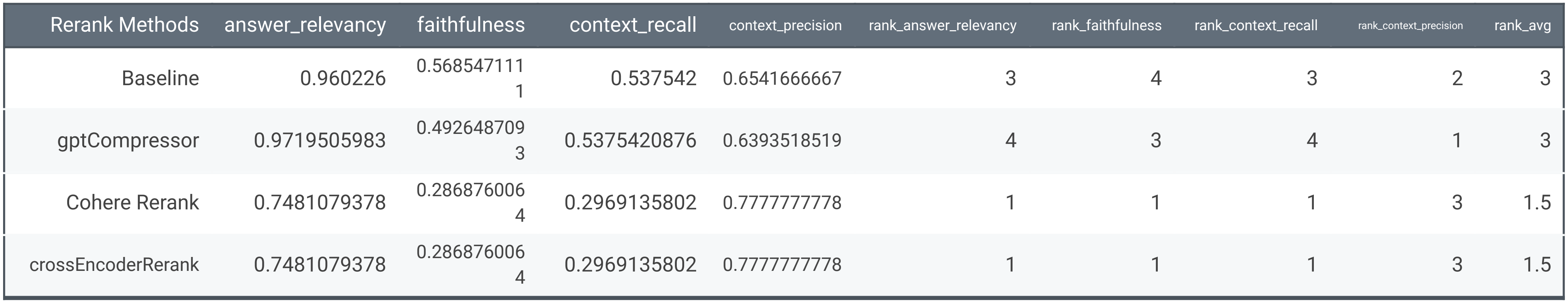
Chunking Strategy	answer_relevancy	faithfulness	context_recall	context_precision	answer_relevancy_rank	faithfulness_rank	context_recall_rank	context_precision_rank	rank_average
0 (Chunk Size: 200, Overlap: 20)	0.9552	0.4109	0.3813	0.6624	2	2	2	3	2.25
1 (Chunk Size: 150, Overlap: 15)	0.6372	0.2407	0.0000	0.0000	1	1	1	1	1.00
2 (Chunk Size: 30, Overlap: 30)	0.9585	0.5833	0.5242	0.6198	3	3	3	2	**2.75**

3. Query Transformation Techniques



Query Transformation Methods	answer_relevancy	faithfulness	context_recall	context_precision	rank_answer_relevancy	rank_faithfulness	rank_context_recall	rank_context_precision	rank_avg
Baseline	0.960226	0.5685471111	0.537542	0.6541666667	3	3	3	3	**3**
Cohere's query translation	0.9470515556	0.6405644444	0.347643	0.6481481111	2	4	1	2	2.25
HyDE	0.866627003	0.5302239183	0.5375420876	0.6541666667	1	1	4	3	2.25
Multi queries	0.9626418889	0.5368481111	0.4517395556	0.3602314444	4	2	2	1	2.25

4. Reranking Techniques



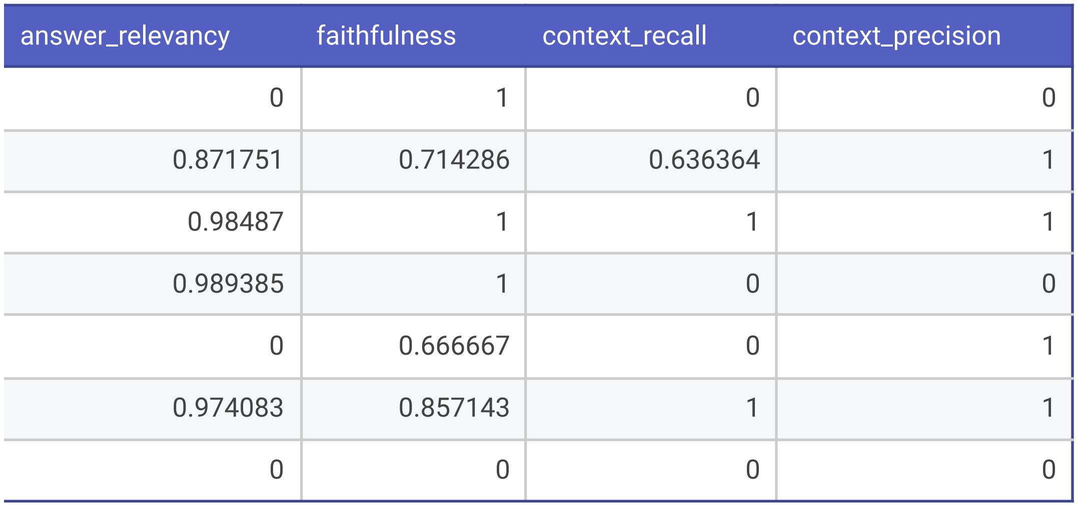
Rerank Methods	answer_relevancy	faithfulness	context_recall	context_precision	rank_answer_relevancy	rank_faithfulness	rank_context_recall	rank_context_precision	rank_avg
Baseline	0.960226	0.5685471111	0.537542	0.6541666667	3	4	3	2	3
gptCompressor	0.9719505983	0.4926487093	0.5375420876	0.6393518519	4	3	4	1	3
Cohere Rerank	0.7481079378	0.2868760064	0.2969135802	0.7777777778	1	1	1	3	1.5
crossEncoderRerank	0.7481079378	0.2868760064	0.2969135802	0.7777777778	1	1	1	3	1.5

RAG-Pipeline Evaluation

Summary

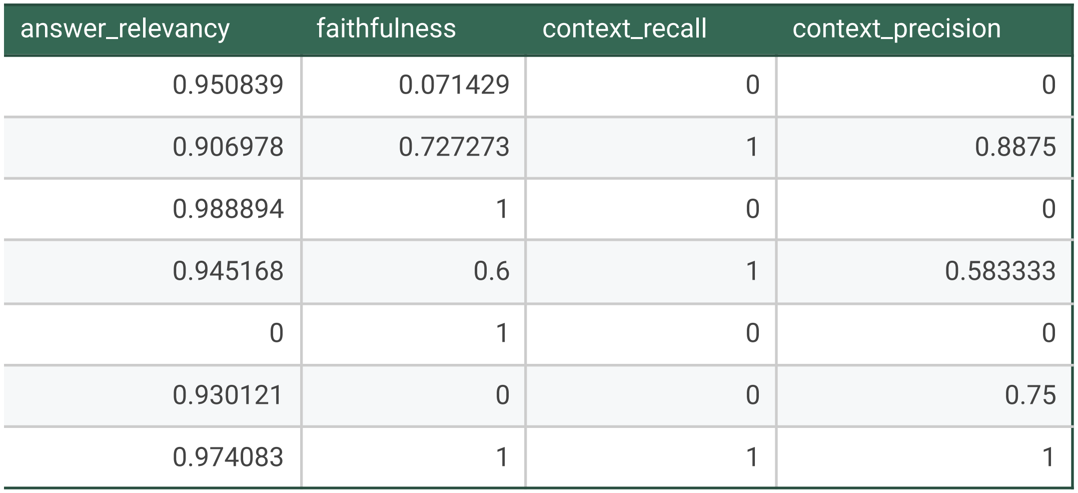
Deployment	avg_answer_relevancy	avg_faithfulness	avg_context_recall	avg_context_precision	rank_answer_relevancy	rank_faithfulness	rank_context_recall	rank_context_precision	rank_avg
OpenAI	0.8137261429	0.628386	0.4285714286	0.4285714286	2	2	2	2	2
Cohere	0.545727	0.7482994286	0.3766234286	0.3766234286	1	3	1	1	1.5
Groq	0.8368711429	0.4527778333	0.7619047143	0.7619047143	3	1	3	3	2.5

1. Cohere Platform



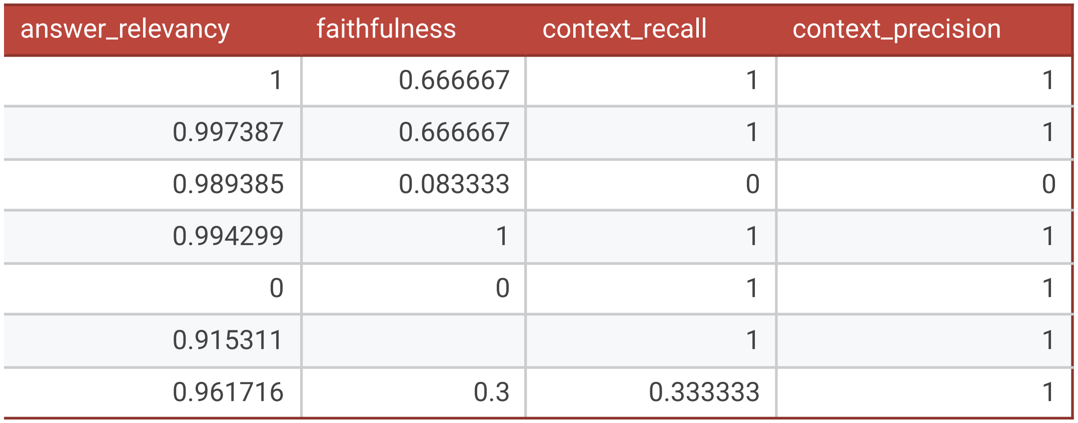
answer_relevancy	faithfulness	context_recall	context_precision
0	1	0	0
0.871751	0.714286	0.636364	1
0.98487	1	1	1
0.989385	1	0	0
0	0.666667	0	1
0.974083	0.857143	1	1
0	0	0	0

2. OpenAI Deployment



answer_relevancy	faithfulness	context_recall	context_precision
0.950839	0.071429	0	0
0.906978	0.727273	1	0.8875
0.988894	1	0	0
0.945168	0.6	1	0.583333
0	1	0	0
0.930121	0	0	0.75
0.974083	1	1	1

3. Groq Deployment



answer_relevancy	faithfulness	context_recall	context_precision
1	0.666667	1	1
0.997387	0.666667	1	1
0.989385	0.083333	0	0
0.994299	1	1	1
0	0	1	1
0.915311		1	1
0.961716	0.3	0.333333	1