📋 Reliability Analysis

🛠️Automatic Evalutating Methods
📐Metrics of RAGAS
📊Test Results

🛠️Automatic Evalutating Methods

Considering that the reliability of our QA system largely depends on the reliability of the relevant documents retrieved by RAG, we designed a method to automatically evaluate the reliability of the five RAG techniques we have studied:

1. First, we need to create question-answer pairs as the ground truth. We have thought about manually writing possible asked questions and answers based on a given URL. However, considering that the URL itself as an input parameter of the entire system is variable, and we will need a large number of samples to produce effective evaluation results, we decided to use LLM's powerful text understanding capabilities to help generate question-answer pairs. Specifically, after we fragment the text obtained through the URL, we randomly select a variable number of fragments, input them to LLM and let it generate possible questions and answers based on these provided texts, and use these question-answer pairs as the ground truth to measure RAG's reliability.
2. Next, we take the generated questions separately as the input of RAG to obtain the retrieved documents and the answers made by LLM based on the retrieved documents of RAG.
3. Finally, we used the a third-party extension ragas and provided the questions, retrieved documents of RAG, answers of LLM based on RAG's retrieved documents and the standard answers as input to ragas to obtain measurement results regarding four metrics: answer relevancy, faithfulness, context precision, context recall.

📐Metrics of RAGAS

- **context precision:** measures the relevance between the retrieved documents and the question, the higher the score is, the less noise(irrelevant information) contained in the retrieved documents.
- **context recall:** measures whether all relevant information required to answer the question have been retrieved, the higher the score is, more relevant information have been retrieved.
- **faithfulness:** measures whether the claims in the generated answer can be referred from the retrieved information,the higher the score is, the more faithful the generated answer is to the retrieved information.
- **answer relevancy:** measures the relevance between the question and the generated answer,the higher the score is, the more relevant the generated answer is to the question.

📊Test Results

The results is tested on 20 question-answer pairs:

RAG Methods	context precision	context recall	faithfulness	answer relevancy
Vectorstore Base Retriever	0.9292	0.9750	0.9833	0.7619
Multiquery	0.8609	0.8250	0.9464	0.8136
Context Compression	0.9347	0.8000	0.8667	0.6755
Parent Document	0.9653	0.8500	0.9896	0.7574

Multivector Summary	0.8625	0.8750	0.9510	0.7222
Multivector Hypothetical	0.8889	0.9083	0.9611	0.7222

As shown in the results, regarding context precision and context recall, all RAG methods can achieve a score higher than 0.80, which indicates a good reliability of the retrieved documents. Also, the faithfulness of all methods are good (higher than 0.85), which means the final answer generated by LLM shows a high relevance to the retrieved text, from which we can conclude that the answer is largely based on the domain-specific knowledge provided by RAG rather than general knowledge of LLM.