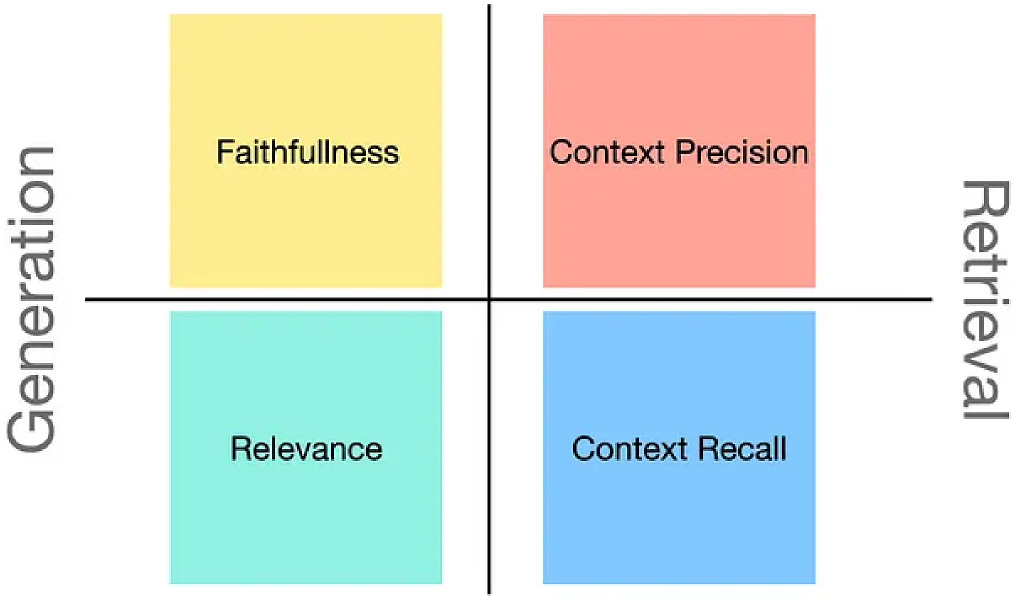
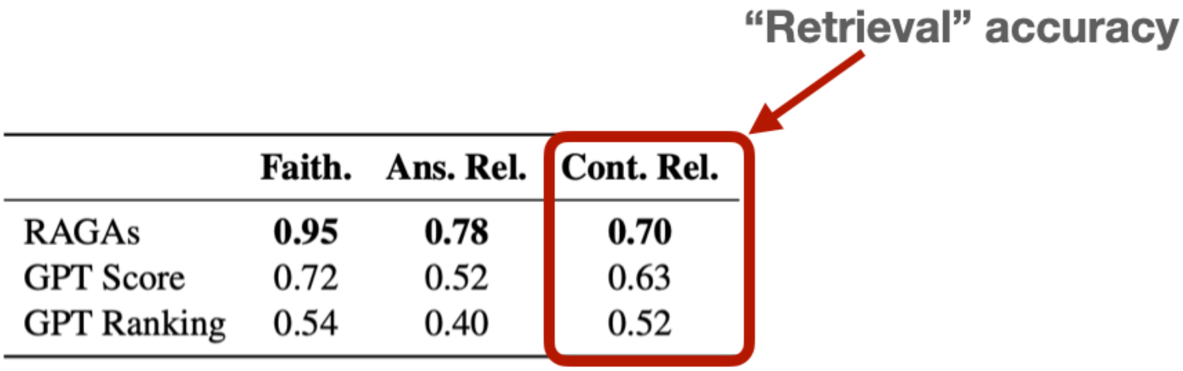
RAGAS :




“Retrieval” accuracy

	Faith.	Ans. Rel.	Cont. Rel.
RAGAs	**0.95**	**0.78**	**0.70**
GPT Score	0.72	0.52	0.63
GPT Ranking	0.54	0.40	0.52

Table 1: Agreement with human annotators in pairwise comparisons of faithfulness, answer relevance and context relevance, using the WikEval dataset (accuracy).

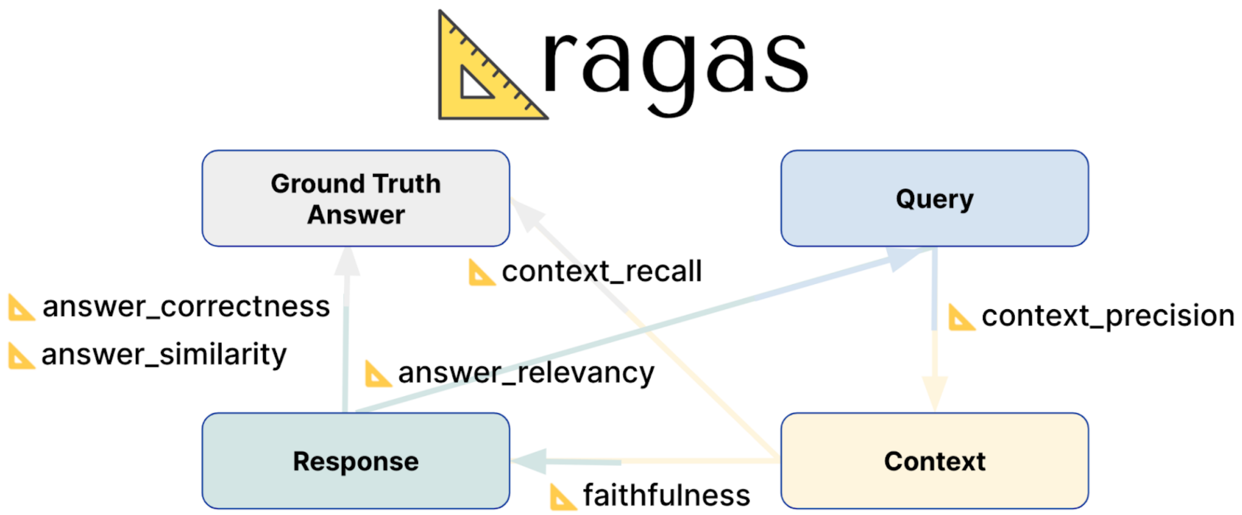
What is Ragas?

Ragas is a specialized evaluation framework designed to assess the performance of Retrieval Augmented Generation (RAG) systems. It provides a structured approach to evaluate the effectiveness of RAG implementations by leveraging advanced Large Language Models (LLMs) as judges. Ragas focuses on automating the evaluation process, offering scalable and cost-effective solutions for assessing AI-generated responses.

Evaluation Data Needed for Ragas

According to the Ragas documentation, your RAG pipeline evaluation will need four key data points.

1. **Question**: The question asked.
2. **Contexts**: Text chunks from your data that best match the question's meaning.
3. **Answer**: Generated answer from your RAG chatbot to the question.
4. **Ground truth answer**: Expected answer to the question.



Ragas Evaluation Metrics

Key Evaluation Metrics

You can find explanations for each evaluation metric, including their underlying formulas, in the documentation. For example, faithfulness. Ragas provides a range of evaluation scores to gauge the effectiveness of RAG systems:

- **Faithfulness**: This score evaluates how accurately the generated answer reflects the information in the provided context. It measures the factual accuracy of the answer, ensuring it aligns with the context from which it is derived. Scores range from 0 to 1, with higher values indicating greater accuracy and consistency.
- **Answer Relevancy**: This answer relevancy metric assesses how well the generated answer responds to the prompt. It focuses on the completeness and relevance of the answer, penalizing incomplete or redundant responses. The relevancy score is derived from the question, context, and answer, with higher scores reflecting better alignment with the prompt.
- **Context Recall**: Context Recall measures how effectively the retrieved context matches the ground truth answer. It calculates the proportion of relevant pieces that were successfully retrieved compared to what was expected. Scores range from 0 to 1, with higher values indicating that a greater portion of relevant context was retrieved.
- **Context Precision**: This metric evaluates whether the most relevant context items are ranked higher than less relevant ones. It checks if all pertinent context chunks appear towards the top of the list. Context Precision is determined using the question, ground truth, and contexts, with higher scores indicating better ranking of relevant information.
- **Context Relevancy**: This context relevance score assesses how relevant the retrieved context is to the question. It measures the degree to which the context matches the intent of the query. The metric ranges from 0 to 1, with higher values showing that the context is more pertinent to the question.
- **Context Entity Recall**: This metric calculates how well the retrieved context captures entities mentioned in the ground truth. It measures the proportion of entities found in both the context and ground truth relative to the total number of entities in the ground truth. Higher scores indicate better capture of important entities in the context.