**Context Recall and Context Precision**

Both **Context Recall** and **Context Precision** are metrics used to evaluate the effectiveness of a document retrieval system. They assess different aspects of how well the system retrieves relevant documents in response to a query.

1. **Context Precision**

* **Definition**: Measures the proportion of retrieved documents that are **relevant** to the query out of all the documents retrieved.

A black text on a white background

Description automatically generated

* Query: "Causes of deforestation."
* Retrieved documents: 5 total, of which 3 are relevant and 2 are irrelevant.
* Precision = 3/5​=0.6 (60%).

1. **Context Recall**

* **Definition**: Measures the proportion of **relevant documents** that are successfully retrieved out of all the relevant documents available in the dataset.

A black text on a white background

Description automatically generated

**Example**:

* Query: "Causes of deforestation."
* Ground truth (relevant documents): 5 documents.
* Retrieved documents: 3
* Recall = 3/5 = 0.6 (60%).

**How They Complement Each Other**

* **High Recall + Low Precision**: Many relevant documents are retrieved, but there’s also a lot of noise (irrelevant documents).
* **High Precision + Low Recall**: Most retrieved documents are relevant, but some important ones are missing.
* **High Recall + High Precision**: Ideal case—comprehensive and accurate retrieval.

In practice:

* **Recall** is prioritized when missing relevant documents is costly (e.g., medical research, legal discovery).
* **Precision** is prioritized when irrelevant information can overwhelm the user (e.g., search engine queries).

**Faithfulness:**

Faithfulness metric measures the factual consistency of the generated answer against the given context.

It is calculated from answer and retrieved context. The answer is scaled to (0,1) range. Higher the better. 1

**Faithfulness score =** (Number of claims in the generated answer that can be inferred from given context) / (Total number of claims in the generated answer)

**User input:**

How many articles and downloadable resources are there in the Selenium WebDriver python course?

**Retrieved contexts:** Complete Understanding on Selenium Python API Methods with real time Scenarios on LIVE Websites\n"Last but not least" you can clear any Interview and can Lead Entire Selenium Python Projects from Design Stage\nThis course includes:\n17.5 hours on-demand video\nAssignments\n23 articles\n9 downloadable resources\nAccess on mobile and TV\nCertificate of completion\nRequirements

**High faithfulness answer:** There are 23 articles and 9 downloadable resources.

**Low faithfulness answer:** There are 23 articles and downloadable resources varies from course

**Response Relevancy**

Response Relevancy metric focuses on assessing how pertinent the generated answer is to the given prompt. A lower score is assigned to answers that are incomplete or contain redundant information and higher scores indicate better relevancy.

Assessment of answer relevance does not consider factuality but instead penalizes cases where the answer lacks completeness or contains redundant details. To calculate this score, the LLM is prompted to generate an appropriate question for the generated answer multiple times, and the mean cosine similarity between these generated questions and the original question is measured.

Example:

Question: Where is France and what is it's capital?

Low relevance answer: France is in western Europe.

High relevance answer: France is in western Europe and Paris is its capital.

**How different it is from Faithfulness?**

Faithfulness metric measures the number of factual claims of the generated answer against the given retrieval context where as Response Relevancy metric focuses on assessing how relevant the generated answer is to the given prompt

Factual Correctness:

Factual Correctness is a metric that compares and evaluates the factual accuracy of the generated response with the reference (ground truth/expected value). This metric is used to determine the extent to which the generated response aligns with the reference. The factual correctness score ranges from 0 to 1, with higher values indicating better performance.

Rubrics based criteria scoring:

This metric that is used to evaluate response. The rubric consists of descriptions for each score, typically ranging from 1 to 10. The response here is evaluation based on score\_descriptions and ground truth.

rubrics = {

"score1\_description": "The response is incorrect, irrelevant, or does not align with the ground truth.",

"score2\_description": "The response partially matches the ground truth but includes significant errors, omissions, or irrelevant information.",

"score3\_description": "The response generally aligns with the ground truth but may lack detail, clarity, or have minor inaccuracies.",

"score4\_description": "The response is mostly accurate and aligns well with the ground truth, with only minor issues or missing details.",

"score5\_description": "The response is fully accurate, aligns completely with the ground truth, and is clear and detailed.",

}

**user\_input="Where is the Eiffel Tower located?",  
response="The Eiffel Tower is located in Europe and it is part of France.",  
reference="The Eiffel Tower is located in Paris.",**

﻿