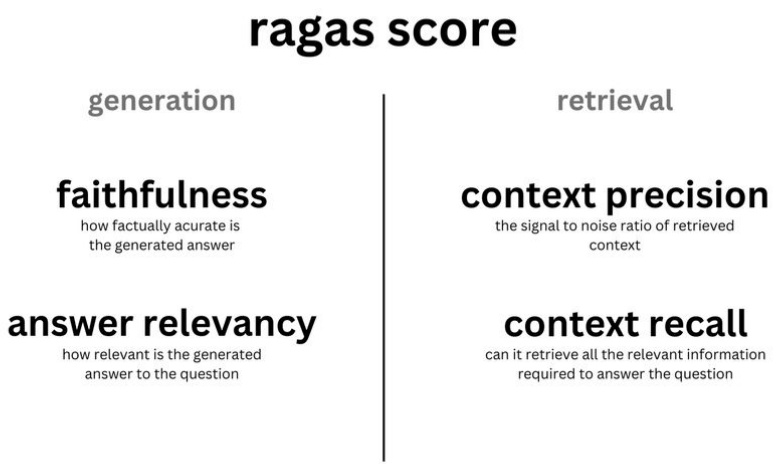
Evaluating LLM and RAG Systems - Metrics

Dipanjan (DJ) Sarkar

Analytics Vidhya

Ragas for evaluating LLM and RAG Systems

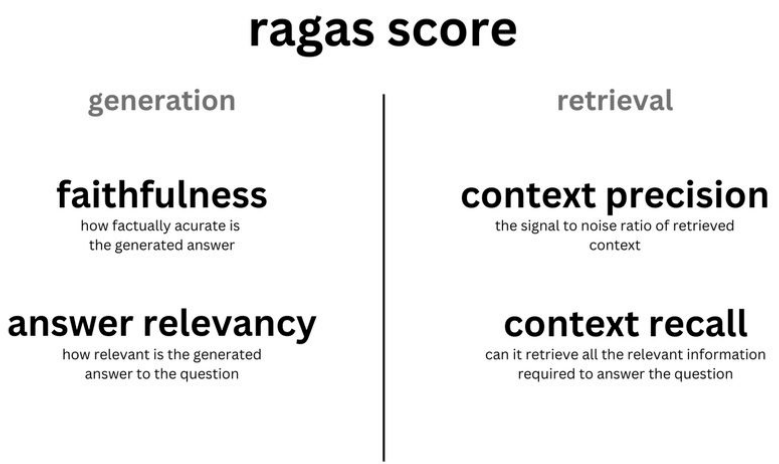


- Ragas is a framework that helps you evaluate your Retrieval Augmented Generation (RAG) pipelines
- Very useful as it provides a suite of metrics to evaluate each stage of a RAG pipeline

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Key Metrics for RAG Evaluation



- Metrics to evaluate specific components of a RAG pipeline include:
  - Faithfulness
  - Answer Relevance
  - Context Precision
  - Context Relevancy
  - Context Recall
  - Context entities recall
- Metrics to evaluate the entire LLM or RAG system include:
  - Answer semantic similarity
  - Answer Correctness

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Faithfulness

$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

Hint

Question: Where and when was Einstein born?

Context: Albert Einstein (born 14 March 1879) was a German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time

High faithfulness answer: Einstein was born in Germany on 14th March 1879.

Low faithfulness answer: Einstein was born in Germany on 20th March 1879.

- Measures the factual consistency of the generated answer against the given context
- Calculated from answer and retrieved context
- Answer is scaled to (0,1) range. Higher the better

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Faithfulness

$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

Hint

Question: Where and when was Einstein born?

Context: Albert Einstein (born 14 March 1879) was a German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time

High faithfulness answer: Einstein was born in Germany on 14th March 1879.

Low faithfulness answer: Einstein was born in Germany on 20th March 1879.

- Generated answer is regarded as faithful if all the claims made in the answer can be inferred from the given context
- Uses the formula as shown on the left

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Faithfulness

$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

Hint

Question: Where and when was Einstein born?

Context: Albert Einstein (born 14 March 1879) was a German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time

High faithfulness answer: Einstein was born in Germany on 14th March 1879.

Low faithfulness answer: Einstein was born in Germany on 20th March 1879.

- Let's examine how faithfulness was calculated using the low faithfulness answer:
  - Step 1: Break the generated answer into individual statements.
    - Statement 1: "Einstein was born in Germany."
    - Statement 2: "Einstein was born on 20th March 1879."
  - Step 2: For each of the generated statements, verify if it can be inferred from the given context.
    - Statement 1: Yes
    - Statement 2: No
  - Step 3: Use the formula depicted above to calculate faithfulness.

$$\text{Faithfulness} = \frac{1}{2} = 0.5$$

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Faithfulness in Ragas

```
from datasets import Dataset
from ragas.metrics import faithfulness
from ragas import evaluate

data_samples = {
    'question': ['When was the first super bowl?',
                 'Who won the most super bowls?'],
    'answer': ['The first superbowl was held on Jan 15, 1967',
               'The most super bowls have been won by The New England Patriots'],

    'contexts' : [['The First AFL–NFL World Championship Game was an American football
game played on January 15, 1967, at the Los Angeles Memorial Coliseum in Los Angeles,'],
    ['The Green Bay Packers...Green Bay, Wisconsin.','The Packers compete...Football
Conference']],
}

dataset = Dataset.from_dict(data_samples)
score = evaluate(dataset,metrics=[faithfulness])
score.to_pandas()
```

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Answer Relevance

The Answer Relevancy is defined as the mean cosine similartiy of the original `question` to a number of artifical questions, which where generated (reverse engineered) based on the `answer`:

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} cos(E_{g_i}, E_o)$$

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} \frac{E_{g_i} \cdot E_o}{\|E_{g_i}\|\|E_o\|}$$

Where:

- E_{g_i} is the embedding of the generated question i.
- E_o is the embedding of the original question.
- N is the number of generated questions, which is 3 default.

- Answer Relevancy, focuses on assessing how pertinent or relevant the generated answer is to the given prompt
- Lower score is assigned to answers that are incomplete or contain redundant information and higher scores indicate better relevancy
- Metric is computed using the question, the context and the answer

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Answer Relevance

The Answer Relevancy is defined as the mean cosine similartiy of the original `question` to a number of artifical questions, which where generated (reverse engineered) based on the `answer`:

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} cos(E_{g_i}, E_o)$$

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} \frac{E_{g_i} \cdot E_o}{\|E_{g_i}\|\|E_o\|}$$

Where:

- E_{g_i} is the embedding of the generated question i.
- E_o is the embedding of the original question.
- N is the number of generated questions, which is 3 default.

- To calculate this score, the LLM is prompted to generate an appropriate question for the generated answer multiple times
- The mean cosine similarity between these generated questions and the original question is measured
- If the generated answer accurately addresses the initial question, the LLM should be able to generate questions from the answer that align with the original question

Source: https://docs.ragas.io/en/stable/index.html

Answer Relevance

Hint

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.

- To calculate the relevance of the answer to the given question, we follow two steps:
  - Step 1: Reverse-engineer 'n' variants of the question from the generated answer using a Large Language Model (LLM). For instance, for the first answer, the LLM might generate the following possible questions:
    - Question 1: "In which part of Europe is France located?"
    - Question 2: "What is the geographical location of France within Europe?"
    - Question 3: "Can you identify the region of Europe where France is situated?"
  - Step 2: Calculate the mean cosine similarity between the generated questions and the actual question.

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Answer Relevance in Ragas

```
from datasets import Dataset
from ragas.metrics import answer_relevancy
from ragas import evaluate

data_samples = {
    'question': ['When was the first super bowl?', 'Who won the most super bowls?'],
    'answer': ['The first superbowl was held on Jan 15, 1967', 'The most super bowls have
been won by The New England Patriots'],
    'contexts' : [['The First AFL–NFL World Championship Game was an American football
game played on January 15, 1967, at the Los Angeles Memorial Coliseum in Los Angeles,'],
    ['The Green Bay Packers...Green Bay, Wisconsin.','The Packers compete...Football
Conference']],
}

dataset = Dataset.from_dict(data_samples)
score = evaluate(dataset,metrics=[answer_relevancy])
score.to_pandas()
```

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context Precision

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

Where K is the total number of chunks in `contexts` and $v_k \in \{0, 1\}$ is the relevance indicator at rank k.

- Metric that evaluates whether all of the ground-truth relevant items present in the contexts are ranked higher or not
- Ideally all the relevant chunks must appear at the top ranks
- This metric is computed using the question, ground_truth and the contexts
- Higher score is better

Source: https://docs.ragas.io/en/stable/index.html

Context Precision

Hint

Question: Where is France and what is it's capital? Ground truth: France is in Western Europe and its capital is Paris.

High context precision: ["France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower", "The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history."]

Low context precision: ["The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and", "France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower",]

- To calculate context precision in the low context precision example:
  - Step 1: For each chunk in retrieved context, check if it is relevant or not relevant to arrive at the ground truth for the given question.
  - Step 2: Calculate precision@k for each chunk in the context.

$$\text{Precision@1} = \frac{0}{1} = 0$$

$$\text{Precision@2} = \frac{1}{2} = 0.5$$

  - Step 3: Calculate the mean of precision@k to arrive at the final context precision score.

$$\text{Context Precision} = \frac{(0+0.5)}{1} = 0.5$$

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context Precision in Ragas

```python
from datasets import Dataset
from ragas.metrics import context_precision
from ragas import evaluate

data_samples = {
    'question': ['When was the first super bowl?', 'Who won the most super bowls?'],
    'answer': ['The first superbowl was held on Jan 15, 1967', 'The most super bowls have been won by The New England Patriots'],
    'contexts' : [['The First AFL–NFL World Championship Game was an American football game played on January 15, 1967, at the Los Angeles Memorial Coliseum in Los Angeles,'],
    ['The Green Bay Packers...Green Bay, Wisconsin.','The Packers compete...Football Conference']],
    'ground_truth': ['The first superbowl was held on January 15, 1967', 'The New England Patriots have won the Super Bowl a record six times']
}

dataset = Dataset.from_dict(data_samples)
score = evaluate(dataset,metrics=[context_precision])
score.to_pandas()
```

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context Relevancy

$$\text{context relevancy} = \frac{|S|}{|\text{Total number of sentences in retrieved context}|}$$

- Metric gauges the relevancy of the retrieved context, calculated based on both the question and contexts
- Retrieved context should exclusively contain essential information to address the provided query
- We initially estimate the value of |S| by identifying sentences within the retrieved context that are relevant for answering the given question and use the formula on the left
- Higher scores indicate better relevancy

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context Relevancy

Hint

Question: What is the capital of France?

High context relevancy: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower.

Low context relevancy: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower. The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history.

- To calculate context relevancy in the given example:
  - Step 1: For any context document find out the number of sentences |S| which are relevant to the question
  - Step 2: Count total sentences in the retrieved context |T|
  - Step 3: Score is |S| / |T|

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context Relevancy in Ragas

```python
from ragas.metrics import ContextRelevancy
context_relevancy = ContextRelevancy()


# Dataset({
#     features: ['question','contexts'],
#     num_rows: 25
# })
dataset: Dataset

results = context_relevancy.score(dataset)
```

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context Recall

The formula for calculating context recall is as follows:

$$\text{context recall} = \frac{|\text{GT sentences that can be attributed to context}|}{|\text{Number of sentences in GT}|}$$

- Measures the extent to which the retrieved context aligns with the annotated answer, treated as the ground truth
- Computed based on the ground truth and the retrieved context
- Each sentence in the ground truth answer is analyzed to determine whether it can be attributed to the retrieved context or not
- Higher Score is better

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context Recall

Hint

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

Ground truth: France is in Western Europe and its capital is Paris.

High context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower.

Low context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history.

- To calculate context recall in the low context recall example:
  - Step 1: Break the ground truth answer into individual statements.
    - Statement 1: "France is in Western Europe."
    - Statement 2: "Its capital is Paris."
  - Step 2: For each of the ground truth statements, verify if it can be attributed to the retrieved context.
    - Statement 1: Yes
    - Statement 2: No
  - Step 3: Use the formula to calculate context recall.

$$\text{context recall} = \frac{1}{2} = 0.5$$

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context Recall in Ragas

```python
from datasets import Dataset
from ragas.metrics import context_recall
from ragas import evaluate

data_samples = {
    'question': ['When was the first super bowl?', 'Who won the most super bowls?'],
    'answer': ['The first superbowl was held on Jan 15, 1967', 'The most super bowls have been won by The New England Patriots'],
    'contexts' : [['The First AFL–NFL World Championship Game was an American football game played on January 15, 1967, at the Los Angeles Memorial Coliseum in Los Angeles,'],
    ['The Green Bay Packers...Green Bay, Wisconsin.','The Packers compete...Football Conference']],
    'ground_truth': ['The first superbowl was held on January 15, 1967', 'The New England Patriots have won the Super Bowl a record six times']
}

dataset = Dataset.from_dict(data_samples)
score = evaluate(dataset,metrics=[context_recall])
score.to_pandas()
```

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context entities Recall

To compute this metric, we use two sets, GE and CE, as set of entities present in `ground_truths` and set of entities present in `contexts` respectively. We then take the number of elements in intersection of these sets and divide it by the number of elements present in the GE, given by the formula:

$$\text{context entity recall} = \frac{|CE \cap GE|}{|GE|}$$

- Metric gives the measure of recall of the retrieved context, based on the number of entities present in both ground_truths and contexts relative to the number of entities present in the ground_truths alone
- Simply, it is a measure of what fraction of entities are recalled from ground_truths
- It can help evaluate the retrieval mechanism for entities, based on comparison with entities present in ground_truths
- Higher Score is better

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context entities Recall

Hint

Ground truth: The Taj Mahal is an ivory-white marble mausoleum on the right bank of the river Yamuna in the Indian city of Agra. It was commissioned in 1631 by the Mughal emperor Shah Jahan to house the tomb of his favorite wife, Mumtaz Mahal.

High entity recall context: The Taj Mahal is a symbol of love and architectural marvel located in Agra, India. It was built by the Mughal emperor Shah Jahan in memory of his beloved wife, Mumtaz Mahal. The structure is renowned for its intricate marble work and beautiful gardens surrounding it.

Low entity recall context: The Taj Mahal is an iconic monument in India. It is a UNESCO World Heritage Site and attracts millions of visitors annually. The intricate carvings and stunning architecture make it a must-visit destination.

- To calculate context entity recall in the given examples:
  - Step 1: Find entities present in the ground truths.
    - Entities in ground truth (GE) - ['Taj Mahal', 'Yamuna', 'Agra', '1631', 'Shah Jahan', 'Mumtaz Mahal']
  - Step 2: Find entities present in the context.
    - Entities in context (CE1) - ['Taj Mahal', 'Agra', 'Shah Jahan', 'Mumtaz Mahal', 'India']
    - Entities in context (CE2) - ['Taj Mahal', 'UNESCO', 'India']
  - Step 3: Use the formula to calculate context entity recall.

$$\text{context entity recall - 1} = \frac{|CE1 \cap GE|}{|GE|} = 4/6 = 0.666$$

$$\text{context entity recall - 2} = \frac{|CE2 \cap GE|}{|GE|} = 1/6 = 0.166$$

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Context entities Recall in Ragas

```
from datasets import Dataset
from ragas.metrics import context_entity_recall
from ragas import evaluate

data_samples = {
    'contexts' : [['The First AFL–NFL World Championship Game was an American football
game played on January 15, 1967, at the Los Angeles Memorial Coliseum in Los Angeles,'],
    ['The Green Bay Packers...Green Bay, Wisconsin.','The Packers compete...Football
Conference']],
    'ground_truth': ['The first superbowl was held on January 15, 1967', 'The New
England Patriots have won the Super Bowl a record six times']
}

dataset = Dataset.from_dict(data_samples)
score = evaluate(dataset,metrics=[context_entity_recall])
score.to_pandas()
```

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Answer semantic similarity

- Answer Semantic Similarity pertains to the assessment of the semantic resemblance between the generated answer and the ground truth
- This evaluation is based on the ground truth and the answer, with values falling within the range of 0 to 1
- A higher score signifies a better alignment between the generated answer and the ground truth
- Ragas utilizes a cross-encoder model to calculate the semantic similarity score using cosine similarity of the text embeddings

Hint

Ground truth: Albert Einstein's theory of relativity revolutionized our understanding of the universe."

High similarity answer: Einstein's groundbreaking theory of relativity transformed our comprehension of the cosmos.

Low similarity answer: Isaac Newton's laws of motion greatly influenced classical physics.

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Answer semantic similarity in Ragas

```
from datasets import Dataset
from ragas.metrics import answer_similarity
from ragas import evaluate


data_samples = {
    'question': ['When was the first super bowl?', 'Who won the most super bowls?'],
    'answer': ['The first superbowl was held on Jan 15, 1967', 'The most super bowls
have been won by The New England Patriots'],
    'ground_truth': ['The first superbowl was held on January 15, 1967', 'The New
England Patriots have won the Super Bowl a record six times']
}

dataset = Dataset.from_dict(data_samples)
score = evaluate(dataset,metrics=[answer_similarity])
score.to_pandas()
```

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Answer Correctness

Factual correctness quantifies the factual overlap between the generated answer and the ground truth answer. This is done using the concepts of:

- TP (True Positive): Facts or statements that are present in both the ground truth and the generated answer.
- FP (False Positive): Facts or statements that are present in the generated answer but not in the ground truth.
- FN (False Negative): Facts or statements that are present in the ground truth but not in the generated answer.

Now, we can use the formula for the F1 score to quantify correctness based on the number of statements in each of these lists:

$$\text{F1 Score} = \frac{|\text{TP}}{(|\text{TP}| + 0.5 \times (|\text{FP}| + |\text{FN}|))}$$

Next, we calculate the semantic similarity between the generated answer and the ground truth. Once we have the semantic similarity, we take a weighted average of the semantic similarity and the factual similarity calculated above to arrive at the final score. You can adjust this weightage by modifying the `weights` parameter.

- Answer Correctness involves gauging the accuracy of the generated answer when compared to the ground truth
- This evaluation relies on the ground truth and the answer, with scores ranging from 0 to 1
- Answer correctness encompasses two critical aspects
  - Factual similarity
  - Semantic similarity between the generated answer and the ground truth
- Final score is a weighted average

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya

Answer Correctness

Hint

Ground truth: Einstein was born in 1879 in Germany.

High answer correctness: In 1879, Einstein was born in Germany.

Low answer correctness: Einstein was born in Spain in 1879.

- It is computed as the weighted sum \ average of factual correctness and the semantic similarity between the given answer and the ground truth
- Factual correctness quantifies the factual overlap between the generated answer and the ground truth answer
- In the second example:
  - TP: [Einstein was born in 1879]
  - FP: [Einstein was born in Spain]
  - FN: [Einstein was born in Germany]
- We then calculate F1 score using the formula mentioned before
- Then we calculate cosine similarity between the generated answer and the ground truth
- Answer correctness is the weighted average of the above two metrics

Source: https://docs.ragas.io/en/stable/index.html

Answer Correctness in Ragas

```
from datasets import Dataset
from ragas.metrics import faithfulness, answer_correctness
from ragas import evaluate

data_samples = {
    'question': ['When was the first super bowl?', 'Who won the most super bowls?'],
    'answer': ['The first superbowl was held on Jan 15, 1967', 'The most super bowls
have been won by The New England Patriots'],
    'ground_truth': ['The first superbowl was held on January 15, 1967', 'The New
England Patriots have won the Super Bowl a record six times']
}

dataset = Dataset.from_dict(data_samples)
score = evaluate(dataset,metrics=[answer_correctness])
score.to_pandas()
```

Source: https://docs.ragas.io/en/stable/index.html

Analytics Vidhya