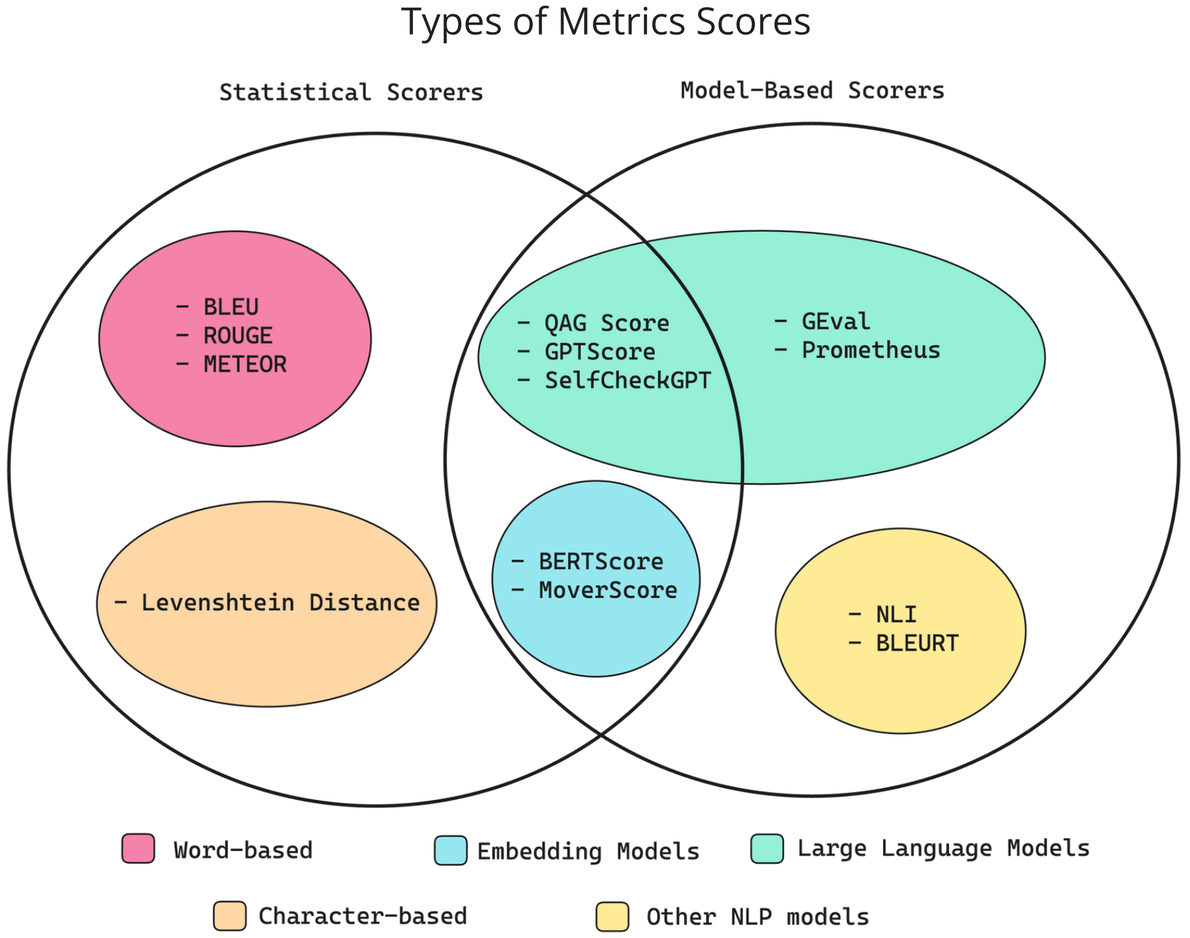
How to evaluate
RAG SYSTEMS

Scope of this tutorial

There are various types of metrics score to evaluate LLM Applications (including RAG Systems).

For this tutorial, the focus is only on **evaluation using Large Language Model** (green circle in the diagram on the right).



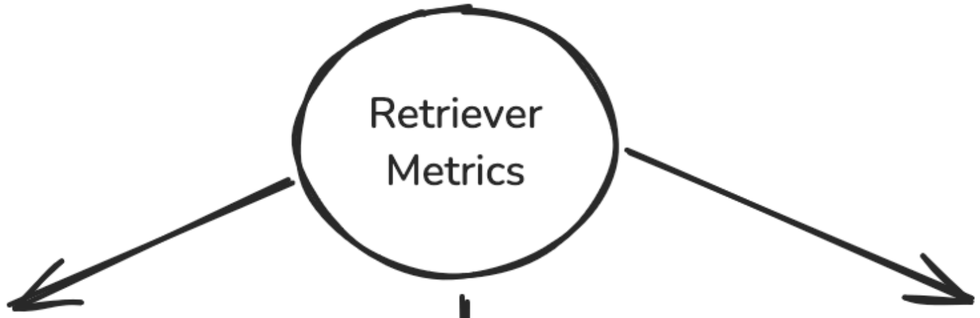
Available Libraries

- **DeepEval** (https://github.com/confident-ai/deepeval)
- Ragas - specifically for RAG (https://github.com/explodinggradients/ragas)
- Prometheus (https://github.com/prometheus/prometheus)
- Phoenix (https://github.com/Arize-ai/phoenix)
- ChainForge (https://github.com/ianarawjo/ChainForge)
- LLM-RAG-Eval (https://github.com/sujitpal/llm-rag-eval)
- TruLens (https://www.trulens.org/)

The rest of the tutorial is based heavily on **DeepEval** library.

By default, DeepEval evaluation works with OpenAI models such as GPT-4o. There is an option to use other custom LLM models, more information here.

Evaluating the Retriever Component



Contextual Precision:
measures your RAG pipeline's retriever by evaluating whether nodes in your retrieval_context that are relevant to the given input are ranked higher than irrelevant ones

Contextual Relevancy:
measures the quality of your RAG pipeline's retriever by evaluating the overall relevance of the information presented in your retrieval_context for a given input

Contextual Recall:
measures the quality of your RAG pipeline's retriever by evaluating the extent of which the retrieval_context aligns with the expected_output.

Contextual Precision

$$\text{Contextual Precision} = \frac{1}{\text{Number of Relevant Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Number of Relevant Nodes Up to Position } k}{k} \times r_k \right)$$

RAG with HIGH Contextual Precision

Input: In what direction does the Sun rise and set?
Retrieved Context:

1. The sun rises in the East.
2. The sun sets in the West.

****

Node 1: Relevant (r1 = 1)
Node 2: Relevant (r2 = 1)

For k=1, Term = 1 x 1 = 1
For k-2, Term = 1 x1 = 1

Sum of terms = 1+1 = 2
Contextual Precision = 2/2 = 1 (the retrieved context are all relevant)

RAG with LOW Contextual Precision

Input: In what direction does the Sun rise and set?
Retrieved Context:

1. The Sun is the centre of our universe.
2. The Sun rises in the East.

****

Node 1: **Not relevant (r1=0)**
Node 2: Relevant (r2 = 1)

For k= 1, Term = 0 x 0 = 0
For k = 2, Term = ½ x 1 = 0.5

Sum of terms = 0 + 0.5 = 0.5
Number of relevant nodes = 1
Contextual Precision = 0.5/1 = 0.5

Contextual Recall

$$\text{Contextual Recall} = \frac{\text{Number of Attributable Statements}}{\text{Total Number of Statements}}$$

RAG with HIGH Contextual Recall

Input: In what direction does the Sun rise and set?
Expected Output: The Sun rises in the East, it sets in the West.
Retrieved Context:

1. The Sun rises in the East.
2. The Sun sets in the West.

****

Statement 1: “The sun rises in the East” is attributed to Node1.

Statement 2: “….sets in the West.” is attributed to Node 2.

Contextual Recall = 2/ 2 = 1

RAG with LOW Contextual Recall

Input: In what direction does the Sun rise and set?
Expected Output: The Sun rises in the East, it sets in the West.
Retrieved Context:

1. The Sun is the centre of our universe.
2. The Sun rises in the East.

****

Statement 1: “The Sun rises in the East” is attributed to Node2.

Statement 2: “….set in the West.” is Not attributable.

Contextual Recall = 1/ 2 = 0.5

Contextual Relevancy

$$\text{Contextual Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

RAG with HIGH Contextual Relevancy

Input: In what direction does the Sun rise and set?
Expected Output: The Sun rises in the East, it sets in the West.
Retrieved Context:
1. The Sun rises in the East.
2. The Sun sets in the West.

****

Node 1: Relevant to the input
Node 2: Relevant to the input

Contextual Relavancy = 2/ 2 = 1

RAG with LOW Contextual Relevancy

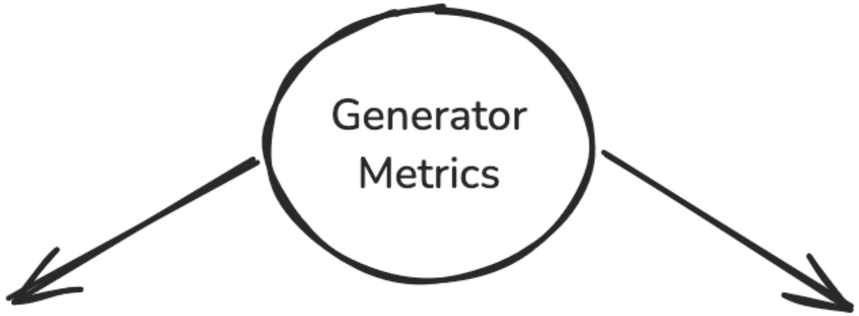
Input: In what direction does the Sun rise and set?
Expected Output: The Sun rises in the East, it sets in the West.
Retrieved Context:
1. The Sun is the centre of our universe.
2. The Sun rises in the East.

****

Node 1: NOT relevant (Question asked about direction of sun rise and sun set and not attributes of the Sun.)
Node 2: Relevant to the input

Contextual Relevancy = 1/ 2 = 0.5

Evaluating the Generator Component



Faithfulness:
measures the quality of your RAG pipeline's generator by evaluating whether the actual_output factually aligns with the contents of your retrieval_context.

Answer Relevancy:
measures the quality of your RAG pipeline's generator by evaluating how relevant the actual_output of your LLM application is compared to the provided input.

Faithfulness

$$\text{Faithfulness} = \frac{\text{Number of Truthful Claims}}{\text{Total Number of Claims}}$$

RAG with HIGH Contextual Relevancy

Input: In what direction does the Sun rise and set?
Retrieved Context:
1. The Sun rises in the East.
2. The Sun sets in the West.

Generated Output: The Sun rises in the East, it sets in the West.

****

Claim 1 in Generated output: Truthful to Node 1
Claim 2 in Generated output: Truthful to Node 2

Number of Truthful claims: 2
Total number of claims: 2

Faithfulness = 2/2 = 1

RAG with HIGH Contextual Relevancy

Input: In what direction does the Sun rise and set?
Retrieved Context:
1. The Sun is the centre of our universe.
2. The Sun rises in the East.

Generated Output: The Sun is the centre of our universe, it rises in the East.

****

Claim 1 in Generated output: Truthful to Node 2
Claim 2 in Generated output: Truthful to Node 1

Number of Truthful claims: 2
Total Number of claims: 2

Faithfulneess = 2/2 = 1

Note: Although the Generated Output is not answering the input question, but it is FAITHFUL to the retrieved context.

Answer Relevancy

$$\text{Answer Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

RAG with HIGH Contextual Relevancy

Input: In what direction does the Sun rise and set?
Generated Output: The Sun rises in the East and sets in the West.

****

Statement 1: “The Sun rises in the East” is Relevant to the input
Statement 2: “….sets in the West.” is Relevant to the input

Number of relevant statement: 2
Total Number of statements: 2

Answer Relavancy = 2/ 2 = 1

RAG with LOW Contextual Relevancy

Input: In what direction does the Sun rise and set?
Generated Output: The Sun is the centre of our universe, it rises in the East.

****

Statement 1: “The Sun is the centre of our universe” is NOT relevant.
Statement 2: “it rises in the East.” is Relevant to the input

Number of relevant statements: 1
Total Number of statements: 2

Answer Relevancy = 1/ 2 = 0.5

References

1. https://medium.com/@med.el.harchaoui/rag-evaluation-metrics-explained-a-complete-guide-dbd7a3b571a8
2. https://docs.confident-ai.com/docs/metrics-answer-relevancy
3. Dipanjan Sarkar's "Comprehensive Guide to LLM & RAG System Evaluation Metrics"
4. https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

Give it a try - sample code below

Colab [Link]

Dr. Meisin Lee

https://www.linkedin.com/in/meisinlee/

https://medium.com/@meisinlee

https://portfolio-meisins-projects.vercel.app/