ROUGE

They check how much the answer overlaps with a perfect answer (n-gram match) but might miss good answers with different wording. Other metrics and methods are better suited for evaluating complex QA systems.
Resource: https://github.com/neural-dialogue-metrics/rouge

Input: questions_and_answers = [(question, answer), (question, answer), …]
results = [generated for question 1, generated for question 2, …]

Code:

```
from evaluate import load

rouge = load('rouge')
scores = rouge.compute(predictions=results, references=[answer for _,
answer in questions_and_answers])
print(scores)
```

MAUVE

MAUVE is a measure the gap between neural text and human text with the eponymous, e.g., how far the text written by a model is the distribution of human text, using samples from both distributions.
MAUVE measure, introduced in this paper (NeurIPS 2021 Ooutstanding Paper).This metrics is a wrapper around the official implementation of MAUVE:
https://github.com/krishnap25/mauve

Resource: https://huggingface.co/spaces/evaluate-metric/mauve

Code:

```
from evaluate import load
rouge = load('mauve')
scores = rouge.compute(predictions=["Q1 generated answer", "Q2 generated
answer" ], references=["Q1 reference answer", "Q2 reference answer"])
print(scores.mauve)
```

RAG Retriever Part Metrics

1) **Precision at K (precision_at_k)**
   Precision at K measures the proportion of relevant items among the top K retrieved items. It focuses on the accuracy of the retrieved results by determining how many of the items retrieved are relevant to the query.

2) **Recall at K (recall_at_k)**
   Recall at K measures the proportion of relevant items that were successfully retrieved out of all the relevant items in the dataset. It assesses the comprehensiveness of the retrieval system by indicating how many relevant items were actually found among the top K retrieved items.

3) **Normalized Discounted Cumulative Gain at 3 (ndcg_at_3)**
   Normalized Discounted Cumulative Gain at 3 is a measure that evaluates the ranking quality of the top 3 retrieved items. It considers both the relevance and the position of the retrieved items in the ranked list, giving higher scores to relevant items that are placed higher in the list.

Resouses: https://mlflow.org/docs/latest/llms/rag/notebooks/retriever-evaluation-tutorial.html

Relevance Metrics

What It Measures: The relevance_metric quantifies how relevant the RAG system's answers are to the input questions. This metric is critical for understanding the accuracy and contextual appropriateness of the system's responses.

We utilize the relevance_metric to evaluate the quality of answers provided by the RAG system. It serves as a quantitative measure of the system's content accuracy, reflecting its capability to generate useful and precise responses.

Resouces: https://mlflow.org/docs/latest/llms/rag/notebooks/mlflow-e2e-evaluation.html

Latency

What It Measures: The latency metric captures the response time of the RAG system. It measures the duration taken by the system to generate an answer after receiving a query.

This evaluation is vital for understanding the system's performance in a production environment, where timely responses are as important as their accuracy.

Resouces: https://mlflow.org/docs/latest/llms/rag/notebooks/mlflow-e2e-evaluation.html

Faithfulness

Faithfulness measures hallucinations or the generation of information not present in the context.

In RAG models, faithfulness assesses how factually consistent the generated response is with the information retrieved from the context source. It essentially measures how well the model adheres to the retrieved content in its response.

Resources:
https://mlflow.org/docs/latest/llms/llm-evaluate/notebooks/rag-evaluation-llama2.html

RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture paper

https://arxiv.org/pdf/2401.08406.pdf

RAG evaluation Metrics Discussed in Article

Recall@top-3:

Measures how many times RAG retrieves the original excerpt given a question.

Formula:

Number of questions where original excerpt is retrieved / Total number of questions

Succinctness:

We created a scoring sheet describing what succinct and verbose answers might contain. No explicit formula is provided, but we prompted LLM with the scoring sheet, the ground truth answer, and the LLM answer, and asked for a grade on a scale from 1 to 5.

Example:

Score: 5

Question: What is the scientific name of the white pine weevil?

Answer: The scientific name of the white pine weevil is Pissodes strobi.

Reference Answer: Pissodes strobi

Explanation: The response is on point and does not contain any additional information. The user is able to understand the point quickly.

Correctness:

We created a scoring sheet describing what a complete, partially correct, or incorrect answer should contain. We prompted GPT-4 with the scoring sheet, the ground truth answer, and the LLM answer, and asked for a grade of correct, incorrect, partially correct.

Diversity:

Evaluates the variety in the generated responses for similar or related queries. This metric is important for understanding if the model can generate different, yet relevant, answers rather than repeating the same responses.

(Diversity can be measured using several approaches, such as calculating the variety of n-grams across responses, but no specific formula is provided.)

Human Evaluation:

Beyond automated metrics, the authors likely incorporate human judgment to assess aspects like naturalness (how human-like the responses are), helpfulness (how useful the information provided is), and overall user satisfaction.

Fine Tuned model evaluation Metrics Discussed in Article

Exact Match (EM):

This metric measures the percentage of responses that exactly match the ground truth answers. It's a strict metric, requiring the generated answer to be identical to the reference answer to count as correct.

- Formula: $\text{EM} = \frac{\text{Number of exact matches}}{\text{Total number of questions}} \times 100$

F1 Score:

The F1 score is used to evaluate the model's accuracy in generating answers by calculating the harmonic mean of precision and recall. Precision measures the relevancy of generated words, while recall assesses how well the model captures all relevant words in the ground truth.

- Formula: $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Knowledgeable F1 (KF1):

A variant of the F1 score that specifically focuses on the accuracy of knowledge-related content in the responses. KF1 evaluates how well the generated answers incorporate and accurately represent specific facts or details from the source material.

While the exact formula for KF1 isn't provided, it can be inferred to adapt the standard F1 formula to specifically assess the presence and accuracy of knowledge-based content.

Unanswerable Question Accuracy:

This metric evaluates the model's ability to correctly identify and appropriately respond to questions that cannot be answered with the available knowledge or due to their nature (e.g., because they are ambiguous or the information is unavailable).

Specific measurement criteria or formula for this metric might involve determining the percentage of correctly handled unanswerable questions out of the total unanswerable questions posed.

Answer Evaluation

Coherence:

Comparison of coherence between ground truths and predictions given the context. The metric provides a score between one to five, where one means that the answer lacks coherence and five means the answer has perfect coherency.

Relevance:

Relevance measures how well the answer addresses the main aspects of the question based on the context. The metric rates from 1 to 5, where 5 means the answer has perfect relevance.

Groundedness:

The metric defines whether the answer follows logically from the information contained in the context or not and provides an integer score to determine how grounded the answer is.