**RAG Evaluation Approaches in RAGEvaluationApp**

The RAGEvaluationApp implements a multi-tiered evaluation framework to assess Retrieval-Augmented Generation (RAG) systems, leveraging three distinct approaches: **RAGAS**, **LLM-based evaluation**, and **ML-based fallback scoring**. These methods work in a cascading hierarchy—RAGAS as the primary evaluator, followed by LLM scoring, and finally ML-based scoring as a fallback—to ensure robust and reliable assessment of RAG performance. This document explains each approach in detail, including their mechanics, implementation in the code, and practical implications.

**1. RAGAS Evaluation**

**Overview**

RAGAS (Retrieval-Augmented Generation Assessment Suite) is a specialized framework for evaluating RAG systems by measuring both retrieval and generation quality. It provides seven key metrics, each scored from 0 to 1, where 1 represents optimal performance. These metrics assess various dimensions of RAG output, from factual consistency to semantic alignment with ground truth.

**Metrics**

1. **Faithfulness**: Measures how factually consistent the generated answer is with the retrieved context.
2. **Answer Relevancy**: Evaluates how well the answer addresses the question.
3. **Context Precision**: Assesses the relevance of retrieved context to the question.
4. **Context Recall**: Checks if the retrieved context contains all necessary information for the ground truth.
5. **Context Entity Recall**: Focuses on the presence of key entities from the ground truth in the context.
6. **Answer Similarity**: Compares the semantic meaning of the answer to the ground truth.
7. **Answer Correctness**: Evaluates factual accuracy and completeness against the ground truth.

**Implementation in Code**

* **Method**: \_evaluate\_with\_ragas
* **Process**:
  1. **Setup**: Configures an AWS Bedrock client with a 120-second timeout and no retries, using ChatBedrock (Nova model) for generation and BedrockEmbeddings (Titan model) for embeddings.
  2. **Data Preparation**: Converts input data (question, context, ground truth, answer) into a Hugging Face Dataset object.
  3. **Evaluation**: Calls ragas.evaluate with the dataset, metrics list, LLM, and embeddings. Results are converted to a pandas DataFrame and mapped to a dictionary with metric scores.
  4. **Fallback**: If any input is empty, returns default scores of 0.5 for all metrics.
* **Key Code**:

python

CollapseWrapCopy

bedrock\_client = boto3.client('bedrock-runtime', config=bedrock\_config)

llm = ChatBedrock(model\_id="amazon.nova-pro-v1:0", client=bedrock\_client)

embeddings = BedrockEmbeddings(model\_id="amazon.titan-embed-text-v2:0", client=bedrock\_client)

data = {"question": [question], "contexts": [[str(context)]], "ground\_truth": [ground\_truth], "answer": [answer]}

hf\_dataset = HFDataset.from\_dict(data)

result = evaluate(dataset=hf\_dataset, metrics=metrics, llm=llm, embeddings=embeddings)

**Mechanics**

* **Faithfulness**: Breaks the answer into statements, verifies each against the context using the LLM, and computes the supported proportion.
* **Answer Relevancy**: Uses embeddings to measure semantic similarity between question and answer, penalizing irrelevance.
* **Context Precision/Recall**: Analyzes context relevance and completeness via LLM reasoning and embeddings.
* **Entity Recall**: Extracts entities from ground truth and checks their presence in context.
* **Similarity/Correctness**: Computes cosine similarity of embeddings and assesses factual alignment.

**Strengths**

* **Comprehensive**: Covers both retrieval (context metrics) and generation (answer metrics).
* **Standardized**: Built on a research-backed framework, ensuring consistency.
* **Granular**: Provides detailed, metric-specific insights.

**Weaknesses**

* **Dependency**: Relies on LLM and embedding quality, which can fail with complex inputs.
* **Resource-Intensive**: Requires API calls to Bedrock, potentially slow or costly.
* **Fragility**: May raise exceptions (e.g., timeouts), necessitating fallbacks.

**Role in the App**

RAGAS is the primary evaluation method, invoked first in \_render\_evaluate\_rag\_page. It offers a robust baseline for scoring RAG performance, with results visualized in the "View Results" page.

**2. LLM-Based Evaluation**

**Overview**

When RAGAS fails (e.g., due to API errors or timeouts), the app falls back to an LLM-based evaluation approach. This method uses an LLM (Amazon-Nova-Pro-v1) to directly score each RAGAS metric by prompting it with specific questions about the input data.

**Metrics**

The same seven RAGAS metrics are evaluated, but scores are derived from LLM responses rather than RAGAS’s internal logic.

**Implementation in Code**

* **Method**: \_evaluate\_with\_llm
* **Process**:
  1. **Setup**: Configures a Bedrock client and instantiates ChatBedrock with the Nova model.
  2. **Prompt Design**: Defines a prompt for each metric, asking the LLM to provide a score from 0 to 1. Example:
     + Faithfulness: "On a scale from 0 to 1, how factually consistent is the answer with the context?"
  3. **Execution**: Invokes the LLM for each prompt, extracts a numeric score from the response using \_extract\_score.
  4. **Fallback**: Returns 0.5 for any metric if the LLM fails or inputs are empty.
* **Key Code**:

python

CollapseWrapCopy

prompts = {

"faithfulness": f"Question: {question}\nContext: {context\_text}\nAnswer: {answer}\nOn a scale from 0 to 1, how factually consistent is the answer with the context? Provide a single number between 0 and 1.",

*# ... other prompts ...*

}

for metric, prompt in prompts.items():

response = llm.invoke(prompt)

scores[metric] = self.\_extract\_score(response.content)

**Mechanics**

* **Prompting**: The LLM interprets the question, context, ground truth, and answer, then assigns a score based on its reasoning.
* **Score Extraction**: Uses regex (re.findall) to find numbers in the response, clamping them to [0, 1]. Defaults to 0.5 if no valid number is found.
* **Context Handling**: Truncates context to 4000 characters to avoid exceeding LLM token limits.

**Strengths**

* **Flexibility**: Adapts to any input without requiring a specialized framework like RAGAS.
* **Resilience**: Works even if RAGAS fails, leveraging the LLM’s general reasoning capabilities.
* **Directness**: Provides a model-driven perspective, potentially capturing nuances missed by automated metrics.

**Weaknesses**

* **Subjectivity**: Scores depend on the LLM’s interpretation, which may vary or be inconsistent.
* **Prompt Sensitivity**: Poorly worded prompts or ambiguous responses can skew results.
* **Resource Use**: Still relies on Bedrock API calls, though less complex than RAGAS.

**Role in the App**

LLM-based evaluation serves as the first fallback in \_render\_evaluate\_rag\_page, triggered if RAGAS raises an exception. It ensures evaluation continuity and provides a secondary layer of insight into model performance.

**3. ML-Based Fallback Scoring**

**Overview**

As a final fallback, the app employs an ML-based approach using text similarity techniques (Jaccard similarity and TF-IDF with cosine similarity) when both RAGAS and LLM evaluation fail. This method is lightweight and deterministic, relying on statistical text analysis rather than LLM reasoning.

**Metrics**

The same seven RAGAS metrics are approximated using similarity scores between pairs of text inputs (question, context, ground truth, answer).

**Implementation in Code**

* **Method**: \_fallback\_scoring
* **Process**:
  1. **Input Validation**: Converts all inputs to strings, handling None values, and returns 0.5 scores if all are empty.
  2. **Jaccard Similarity**:
     + Computes word overlap between two strings (e.g., answer vs. context) as a ratio of intersection to union.
     + Used as a standalone metric if TF-IDF fails or fewer than two valid texts are available.
  3. **TF-IDF and Cosine Similarity**:
     + Vectorizes texts (question, answer, ground truth, context) using TfidfVectorizer.
     + Calculates cosine similarity between vectors to measure semantic overlap.
  4. **Scoring**:
     + Maps similarity scores to metrics (e.g., answer vs. context for faithfulness, answer vs. ground truth for answer\_similarity).
     + Combines TF-IDF and Jaccard for answer\_correctness.
  5. **Fallback**: Uses Jaccard-only scoring if TF-IDF fails, ensuring some result is returned.
* **Key Code**:

python

CollapseWrapCopy

def jaccard\_similarity(str1, str2):

set1, set2 = set(str1.lower().split()), set(str2.lower().split())

union\_len = len(set1 | set2)

return len(set1 & set2) / union\_len if union\_len > 0 else 0.0

vectorizer = TfidfVectorizer()

tfidf\_matrix = vectorizer.fit\_transform([question, answer, ground\_truth, context])

cosine\_sim = cosine\_similarity(tfidf\_matrix, tfidf\_matrix)

scores = {

"faithfulness": cosine\_sim[1, 3], *# Answer vs Context*

"answer\_relevancy": cosine\_sim[1, 0], *# Answer vs Question*

*# ... other mappings ...*

}

**Mechanics**

* **Jaccard Similarity**: Measures lexical overlap (e.g., shared words), simple but effective for basic alignment.
* **TF-IDF**: Weights words by importance (term frequency-inverse document frequency), capturing semantic relevance beyond raw overlap.
* **Cosine Similarity**: Computes the angle between TF-IDF vectors, indicating how similar two texts are in meaning.
* **Metric Approximation**: Assigns similarity scores to RAGAS metrics based on logical pairings (e.g., context vs. ground truth for context\_precision).

**Strengths**

* **Deterministic**: No reliance on external APIs or LLM variability, ensuring consistent results.
* **Lightweight**: Runs locally with minimal computational overhead.
* **Fallback Reliability**: Guarantees scores even in complete failure scenarios.

**Weaknesses**

* **Simplistic**: Lacks the nuanced reasoning of RAGAS or LLM methods, missing context or intent.
* **Lexical Bias**: Focuses on word overlap, potentially misjudging semantic differences (e.g., synonyms).
* **Limited Scope**: Approximates rather than directly measures complex metrics like faithfulness.

**Role in the App**

ML-based scoring is the last resort in \_render\_evaluate\_rag\_page, used when both RAGAS and LLM evaluation fail. It ensures the app always provides some evaluation output, albeit less sophisticated.

**Comparative Analysis**

| **Aspect** | **RAGAS** | **LLM-Based** | **ML-Based** |
| --- | --- | --- | --- |
| **Primary Use** | Main evaluation | First fallback | Final fallback |
| **Metrics** | 7 RAGAS metrics | Same 7, via prompts | Same 7, via similarity |
| **Dependency** | Bedrock LLM + embeddings | Bedrock LLM | None (local ML) |
| **Complexity** | High (framework-based) | Medium (prompt-based) | Low (statistical) |
| **Accuracy** | High (research-backed) | Moderate (subjective) | Low (approximate) |
| **Speed** | Slow (API calls) | Moderate (fewer calls) | Fast (local) |
| **Robustness** | Fragile (API-dependent) | More resilient | Very robust |

**Integration in the Workflow**

**Evaluation Cascade**

* **Step 1**: \_evaluate\_with\_ragas attempts RAGAS evaluation.
  + Success: Returns detailed scores.
  + Failure: Raises exception, triggers LLM fallback.
* **Step 2**: \_evaluate\_with\_llm uses Nova to score metrics.
  + Success: Returns LLM-assigned scores.
  + Failure: Raises exception, triggers ML fallback.
* **Step 3**: \_fallback\_scoring computes similarity-based scores.
  + Always succeeds, providing a baseline.

**Code Flow**

* **Trigger**: Initiated in \_render\_evaluate\_rag\_page when "Evaluate with RAGAS" is clicked.
* **Error Handling**: Wrapped in try-except blocks, with warnings displayed for each fallback transition.
* **Output**: Results are stored in evaluation\_results and saved as a CSV (e.g., RAG\_Results\_YYYYMMDD\_HHMMSS.csv).

**Practical Example**

* **Input**:
  + Question: "What are the 2023 CCAR capital requirements?"
  + Context: "In 2023, CCAR rules set Tier 1 capital at 4.5%."
  + Ground Truth: "2023 CCAR requires a 4.5% Tier 1 capital ratio."
  + Answer: "CCAR 2023 mandates 4.5% Tier 1 capital."
* **RAGAS**: Scores 1.0 for faithfulness, answer\_similarity, etc., using embeddings and LLM checks.
* **LLM Fallback**: Nova rates faithfulness as 1.0 via prompt response: "1.0".
* **ML Fallback**: Cosine similarity between answer and context yields 0.95 for faithfulness.

**Conclusion**

The RAGEvaluationApp employs a tiered evaluation strategy to balance accuracy, robustness, and resilience:

* **RAGAS** provides a gold standard for detailed, framework-based assessment.
* **LLM-Based Evaluation** offers a flexible, reasoning-driven alternative.
* **ML-Based Scoring** ensures a fail-safe, lightweight backup.

Together, these approaches enable comprehensive analysis of RAG systems, from retrieval quality to model performance, supporting applications like CCAR compliance with actionable insights. Users can refine their RAG pipeline by identifying weaknesses (e.g., low context\_recall for retrieval tuning) and comparing models (e.g., Claude vs. Nova) through the app’s visualizations.