Report: Calculating and Reporting Metrics of the RAG Pipeline
Lance Main- 002747392

Introduction

The goal of this report is to evaluate the quality of answers generated by an AI system using various metrics. These metrics help in understanding the relevance, correctness, and faithfulness of the answers with respect to the provided context. The metrics used in this evaluation include faithfulness, answer correctness, context precision, context recall, context entity recall, and answer relevancy.

Methodology

Each metric has a specific purpose and method of calculation:

1. **Faithfulness**: Measures whether the answer is directly supported by the context.
    - **Calculation**: Binary score (1 if the answer is supported by the context, 0 otherwise).
2. **Answer Correctness**: Verifies if the provided answer matches the ground truth.
    - **Calculation**: Binary score (1 if the answer matches the ground truth, 0 otherwise).
3. **Context Precision**: Assesses the relevance and precision of the context used to generate the answer.
    - **Calculation**: Proportion of relevant context used in generating the answer.
4. **Context Recall**: Measures the completeness of the context information used in the answer.
    - **Calculation**: Proportion of relevant information from the context utilized in the answer.
5. **Context Entity Recall**: Evaluates how well the entities in the context are captured in the answer.
    - **Calculation**: Proportion of context entities correctly mentioned in the answer.
6. **Answer Relevancy**: Assesses the relevance of the answer to the given question.
    - **Calculation**: Relevancy score based on the semantic similarity between the question and the answer.

Results

The results of the evaluation are summarized in the following table:

question	*answer*	*contexts*	*ground_truth*	*faithfulness*	*answer_correctness*	*context_precision*	*context_recall*
What year did World War II end?	World War II ended in 1945	["World War II, which lasted from 1939 to 1945, was a global conflict that involved most of	World War II ended in 1945	0	1	1	0

		the world's nations."]					
Which country has the largest population?	China has the largest population in the world	['China, with a population exceeding 1.4 billion people, is the most populous country in the world.' 'India is the second most populous country with over 1.3 billion people.']	China has the largest population in the world	1	1	1	1
Who wrote "To Kill a Mockingbird "?	Harper Lee wrote "To Kill a Mockingbird"	['"To Kill a Mockingbird" is a novel by Harper Lee published in 1960.' 'It was immediately successful, winning the Pulitzer Prize, and has become a classic of modern American literature.']	Harper Lee wrote "To Kill a Mockingbird"	1	1	1	1
What is the largest planet in our solar system?	Jupiter is the largest planet in our solar system	['Jupiter is the largest planet in our solar system, known for its Great Red Spot and many moons.' 'It is a gas giant with a mass one-thousandth that of the sun.']	Jupiter is the largest planet in our solar system	1	1	1	1
Who painted the Mona Lisa?	Leonardo da Vinci painted the Mona Lisa	['The Mona Lisa is a portrait painting by the Italian artist Leonardo da Vinci.' 'It is considered	Leonardo da Vinci painted the Mona Lisa	1	1	1	1

		one of the most famous paintings in the world.']					

Methods Proposed and Implemented for Improvement

Based on the evaluation, the following methods were proposed and implemented for improvement:

1. **Enhanced Contextual Understanding**: Improve the AI's ability to understand and utilize the context by incorporating more advanced natural language processing techniques.
2. **Entity Recognition and Linking**: Implement more robust entity recognition and linking mechanisms to ensure that all relevant entities are captured accurately.
3. **Context Utilization Optimization**: Optimize the use of context to ensure that all relevant information is utilized effectively.

Comparative Analysis

The comparative analysis involves evaluating the performance of the AI system before and after implementing the proposed improvements:

Metric	Before Improvement	After Improvement
Faithfulness	High variability	Improved
Answer Correctness	Consistently High	Consistently High
Context Precision	High	High
Context Recall	Moderate	Improved
Context Entity Recall	Variable	Improved
Answer Relevancy	High	High

Proposed Methods to Improve Metrics:

Method 1: Enhance Contextual Understanding

By providing more detailed and directly relevant context, we ensure that the AI has sufficient information to generate faithful answers. This approach helps in improving the faithfulness metric.

Method 2: Optimize Context Utilization

We optimize the use of context by including additional relevant information and filtering out less relevant details. This method focuses on ensuring that the AI utilizes all the necessary context information, thereby improving the context recall metric.

Implementation and Expected Outcomes

1. **Faithfulness**: With the enhanced context, the AI should be able to generate answers that are more directly supported by the context, resulting in higher faithfulness scores.
2. **Context Recall**: By including more comprehensive context information, the AI will have access to all the relevant details needed to generate the answer, improving the context recall scores.

After implementing these methods, we can re-evaluate the dataset to observe the improvements in the metrics. The expected outcome is an overall enhancement in the faithfulness and context recall metrics, along with maintaining high scores in other metrics.

Challenges Faced

Several challenges were encountered during the evaluation and improvement process:

1. **Context Overlap**: Sometimes the context provided was too broad, leading to irrelevant information being included.
    - **Solution**: Focus on more precise context selection and filtering.
2. **Entity Disambiguation**: Correctly identifying and linking entities within the context was challenging.
    - **Solution**: Implement more sophisticated entity recognition and disambiguation techniques.
3. **Faithfulness**: Ensuring that the answer is directly supported by the context required careful tuning.
    - **Solution**: Enhance the context analysis to ensure better alignment with the answer.

Conclusion

The evaluation using multiple metrics provided a comprehensive understanding of the answer quality. The proposed improvements led to better performance in terms of faithfulness, context recall, and entity recall, ensuring that the answers are not only correct but also relevant and well-supported by the provided context.