Performance Evaluation of Retrieval-Augmented Generation (RAG) Pipeline

Sanjay Bhaskar Kashyap

July 28, 2024

1 Introduction

This report evaluates the performance of the Retrieval-Augmented Generation (RAG) pipeline, focusing on both retrieval and generation metrics. The metrics are calculated based on the provided dataset and the answers generated by the RAG system.

2 Performance Metrics Calculation

2.1 Retrieval Metrics

- **Context Precision:** Measure how accurately the retrieved context matches the user's query.
- **Context Recall:** Evaluate the ability to retrieve all relevant contexts for the user's query.
- **Context Relevance:** Assess the relevance of the retrieved context to the user's query.
- **Context Entity Recall:** Determine the ability to recall relevant entities within the context.
- **Noise Robustness:** Test the system's ability to handle noisy or irrelevant inputs.

2.2 Generation Metrics

- **Faithfulness:** Measure the accuracy and reliability of the generated answers.
- **Answer Relevance:** Evaluate the relevance of the generated answers to the user's query.
- **Information Integration:** Assess the ability to integrate and present information cohesively.
- **Counterfactual Robustness:** Test the robustness of the system against counterfactual or contradictory queries.
- **Negative Rejection:** Measure the system's ability to reject and handle negative or inappropriate queries.
- **Latency:** Measure the response time of the system from receiving a query to delivering an answer.

3 Methods to Improve Metrics

3.1 Proposed Improvements

To enhance the performance of the RAG pipeline, the following methods were proposed and implemented:

- Improvement of Context Precision and Recall by optimizing the retriever settings.
- Enhancement of Faithfulness and Answer Relevance by fine-tuning the generative model.

3.2 Implementation and Impact Analysis

- **Context Precision and Recall:** Adjusted retriever parameters to improve accuracy. The impact analysis shows a significant increase in both precision and recall.
- **Faithfulness and Answer Relevance:** Fine-tuned the generative model using additional training data, resulting in more accurate and relevant answers.

4 Results

Question	Expected Answer
What's the issue reported by Al Grover?	There's a leak coming through his kitchen ceiling.
What is Charlie Johnson's problem?	He received the wrong color scarf.
Why is Ron calling Regency limousine?	His limo is 30 minutes late.
What did Steve Simmons want to do with the blender?	He wants to return a recalled blender and get a different mo
Why is Jeff Matthews complaining?	His neighbors are being loud and disturbing his newborn bab

5 Comparative Analysis

5.1 Performance Before Improvements

The initial evaluation of the RAG pipeline showed moderate performance in most metrics, with room for improvement in precision and faithfulness.

5.2 Performance After Improvements

After implementing the proposed improvements, there was a noticeable enhancement in both retrieval and generation metrics. Precision and faithfulness saw the most significant improvements, resulting in more accurate and relevant answers.

6 Challenges and Solutions

6.1 Challenges

- Difficulty in optimizing retriever settings to balance precision and recall.
- Ensuring the generative model remains faithful to the context while generating relevant answers.

6.2 Solutions

- Iterative testing and fine-tuning of retriever parameters.
- Incorporating more diverse training data to improve the generative model's accuracy.

7 Conclusion

The evaluation and improvement of the RAG pipeline demonstrate the effectiveness of targeted optimizations. By enhancing precision and faithfulness, the system now provides more accurate and relevant answers, improving the overall user experience.