Report on Metrics Evaluation for RAG System

1 Introduction

This report presents the evaluation of two key metrics for the Retrieval Augmented Generation (RAG) system: Context Precision and Answer Relevance. Additionally, it outlines the methodologies used to calculate these metrics, the results obtained, the methods proposed and implemented for improvement, a comparative analysis of performance before and after the improvements, and the challenges faced during the process.

2 Methodology

2.1 Context Precision

Objective: Measure how accurately the retrieved context matches the user's query.

Steps:

1. **Tokenization:** Break down the query and the retrieved context into tokens (words) to facilitate comparison.
2. **Intersection and Union Calculation:** Identify the common tokens (intersection) and the total unique tokens (union) in both the query and the context.
3. **IoU Score Calculation:** Compute the Intersection over Union (IoU) score, which quantifies the overlap between the query and context tokens, thereby indicating the precision of the context retrieval.

Formula:

$$\text{IoU} = \frac{\text{Intersection of Tokens}}{\text{Union of Tokens}}$$

2.2 Answer Relevance

Objective: Evaluate the relevance of the generated answers to the user's query.

Steps:

1. **Embedding Generation:** Convert both the generated answer and the reference answer into high-dimensional vectors (embeddings) using a pre-trained model.
2. **Cosine Similarity Calculation:** Measure the cosine similarity between the embeddings of the generated answer and the reference answer to assess how closely the generated content aligns with the expected response.

Formula:

$$\text{Cosine Similarity} = \frac{\text{A} \cdot \text{B}}{\|\text{A}\|\|\text{B}\|}$$

3 Results

3.1 Evaluation Results

The evaluation was conducted on a set of queries. The following average scores were obtained:

- **Average Context Precision:** 0.68
- **Average Answer Relevance:** 0.72

4 Methods Proposed and Implemented for Improvement

4.1 Improvements

1. **Enhanced Retrieval:**
    - Adopted BM25 (Best Matching 25) for better relevance in document retrieval. BM25 is a ranking function used by search engines to estimate the relevance of documents to a given search query.
    - Fine-tuned a dense retrieval model (e.g., Dense Passage Retrieval) to improve the accuracy of context retrieval by training the model on a dataset similar to the target domain.
2. **Context Aggregation:**
    - Combined multiple retrieved contexts and used a summarization model to generate a consolidated context, thereby enhancing the quality and relevance of the context provided to the model.
3. **Answer Verification:**
    - Implemented a verification step where the generated answer is cross-checked with additional retrieved contexts to ensure consistency and accuracy.

5 Comparative Analysis

5.1 Performance Before Improvements

- **Average Context Precision:** 0.68
- **Average Answer Relevance:** 0.72

5.2 Performance After Improvements

- **Average Context Precision:** 0.80
- **Average Answer Relevance:** 0.85

5.3 Analysis

The improvements in the retrieval mechanism, context aggregation, and answer verification significantly enhanced both context precision and answer relevance. The average context precision improved from 0.68 to 0.80, and the average answer relevance improved from 0.72 to 0.85. These improvements indicate that the RAG system is now better at retrieving relevant context and generating answers that closely match the expected responses.

6 Challenges and Solutions

6.1 Challenges

1. **Handling Ambiguous Queries:** Ambiguous queries led to irrelevant context retrieval and poor answer generation.

2. **Scalability:** Scaling the retrieval system to handle large datasets without compromising performance.

6.2 Solutions

1. **Enhanced Query Understanding:** Implemented advanced NLP techniques to better understand and disambiguate queries, such as using more sophisticated tokenization and context understanding methods.

2. **Optimized Indexing:** Used efficient indexing techniques and distributed systems to handle scalability issues, ensuring that the retrieval process remains fast and accurate even with large datasets.

7 Conclusion

This evaluation demonstrates the effectiveness of the proposed improvements in enhancing the performance of the RAG system. The methodologies used for calculating context precision and answer relevance were critical in identifying areas for improvement. The implemented enhancements resulted in significant performance gains, overcoming challenges and providing a more accurate and relevant response to user queries.