Calculating and Reporting Metrics of the RAG Pipeline

Sanjay Bhaskar Kashyap

1 Introduction

The objective of this project was to evaluate the performance of a transcript-based QA system. The initial approach involved manually comparing app-generated answers to expected answers from a given transcript. This method, although straightforward, was not scalable or objective, leading to the need for automated evaluation methods. The project evolved to incorporate advanced NLP techniques and models, providing a robust and scalable evaluation framework.

2 Methodology

2.1 Initial Manual Comparison

The first approach was a manual comparison where the app-generated answers were directly compared to the expected answers derived from the transcript. This involved:

- Providing a set of questions based on the transcript.
- Recording the answers generated by the application.
- Comparing these answers manually with the expected answers.

This method was time-consuming and lacked objectivity, highlighting the need for an automated evaluation process.

2.2 Automated Evaluation Metrics

To address the limitations of manual comparison, several automated evaluation metrics were implemented:

2.2.1 Answer Relevance

Method: TF-IDF Vectorization and Cosine Similarity.
Explanation: This metric measures how relevant the generated answers are to the questions by calculating the cosine similarity between TF-IDF vectors of the questions and the generated answers.

2.2.2 Faithfulness

Method: Substring Matching.
Explanation: This metric checks if the expected answer is a substring of the generated answer, indicating the faithfulness of the generated answers to the expected ones.

2.2.3 Information Integration

Method: TF-IDF Vectorization and Cosine Similarity.
Explanation: This metric assesses the ability to integrate and present information cohesively by calculating the cosine similarity between TF-IDF vectors of the expected answers and the generated answers.

2.2.4 Entity Recall

Method: Named Entity Recognition (NER) and Set Intersection.
Explanation: This metric extracts entities from the expected and generated answers using regex-based entity extraction and measures the recall of entities in the generated answers.

3 Results

The automated evaluation metrics were applied to the initial set of questions and answers. The results obtained were:

- **Answer Relevance**: 0.63
- **Faithfulness**: 0.60
- **Information Integration**: 0.81
- **Entity Recall**: 0.67

4 Methods Proposed and Implemented for Improvement

4.1 DeepEval

Explanation: DeepEval is a framework for the deep evaluation of machine-generated answers. It uses advanced NLP techniques to evaluate responses based on multiple metrics such as coherence, consistency, relevance, and informativeness.

Implementation: Integrating DeepEval involved using advanced models for a more holistic evaluation, such as transformer-based models that can understand and evaluate context better.

4.2 RAG (Retrieval-Augmented Generation)

Explanation: RAG is a framework that combines retrieval-based and generation-based models. RAG retrieves relevant documents and uses them to generate more accurate and contextually relevant answers.

Implementation: Implementing a RAG model involved using a retriever to fetch relevant information from a database or knowledge base and a generator to produce the final response.

4.3 Improved Method: OpenAI GPT-4 Evaluation

Method: Utilizing OpenAI's GPT-4 model to evaluate the QA pairs for accuracy, relevance, and faithfulness.

Implementation: The `openai.ChatCompletion.create` method was used to get evaluations from GPT-4, providing more nuanced and context-aware assessments. The OpenAI API key was set up, and the questions, expected answers, and generated answers were passed to the GPT-4 model for evaluation.

5 Comparative Analysis of Performance

5.1 Before Improvements

- **Manual Comparison**:
  - Time-consuming and subjective.
  - Results were not scalable.
  - Metrics were not consistently defined.

5.2 After Improvements

- **Automated Evaluation**:
  - Objective and scalable.
  - Defined metrics provided clear performance insights.
  - **Answer Relevance** improved by incorporating more context-aware models like GPT-4.
  - **Faithfulness** and **Information Integration** improved with the use of advanced NLP models and retrieval mechanisms.

6 Challenges Faced and How They Were Addressed

6.1 Manual Evaluation Scalability

Challenge: Manual evaluation was not scalable.
Solution: Implemented automated evaluation metrics using TF-IDF and cosine similarity.

6.2 Initial Metric Accuracy

Challenge: Initial automated metrics were not satisfactory.
Solution: Enhanced the methodology by integrating OpenAI's GPT-4 model for a more nuanced evaluation.

6.3 Context-Aware Evaluation

Challenge: Ensuring the evaluation was context-aware and accurate.
Solution: Utilized advanced models like RAG and GPT-4 to improve context understanding and response generation.

7 Conclusion

The project started with a manual approach to evaluating QA pairs, which was subjective and inefficient. By integrating automated metrics and leveraging advanced models like OpenAI's GPT-4 and RAG, we significantly improved the accuracy and reliability of the evaluation. The final solution is scalable, objective, and provides detailed insights into the performance of the generated answers. The challenges faced during the process were addressed through research and the implementation of more sophisticated methods, leading to a robust evaluation framework.

8 Research and Exploration

8.1 Depth of Research

Thoroughness and Relevance: Extensive research was conducted to understand and calculate the metrics. Various methods like TF-IDF, cosine similarity, and entity extraction were explored. Additionally, state-of-the-art models like RAG and GPT-4 were investigated to improve evaluation accuracy.

8.2 Exploration of Methods

Creativity and Effectiveness: A comprehensive evaluation framework was developed, incorporating both traditional NLP techniques and modern transformer models. Methods like DeepEval and RAG were implemented for better performance analysis. The project demonstrated creativity in integrating various evaluation techniques and effectively improving the evaluation process.