Report on RAG-Based Chatbot Performance Evaluation

1. Introduction

This report presents the evaluation of the RAG (Retrieval-Augmented Generation) pipeline-based chatbot and focuses on key performance metrics. The aim is to understand the effectiveness of the current RAG pipeline and implement improvements to enhance its performance.

2. Performance Metrics Calculation

2.1 Retrieval Metrics

1. **Context Precision**
    - **Definition**: Measures how accurately the retrieved context matches the user's query.
    - **Methodology**: Compare the retrieved contexts with a set of ground truth contexts and calculate the ratio of relevant contexts to the total retrieved contexts.
2. **Context Recall**
    - **Definition**: Evaluates the ability to retrieve all relevant contexts for the user's query.
    - **Methodology**: Calculate the ratio of relevant retrieved contexts to the total number of relevant contexts in the ground truth.
3. **Context Relevance**
    - **Definition**: Assesses the relevance of the retrieved context to the user's query.
    - **Methodology**: Use a relevance scoring system to evaluate the match between retrieved contexts and the query.
4. **Context Entity Recall**
    - **Definition**: Determines the ability to recall relevant entities within the context.
    - **Methodology**: Measure the recall of entities mentioned in the query within the retrieved contexts.
5. **Noise Robustness**
    - **Definition**: Tests the system's ability to handle noisy or irrelevant inputs.
    - **Methodology**: Introduce noise in the query and measure the impact on retrieval accuracy and relevance.

2.2 Generation Metrics

1. **Faithfulness**
    - **Definition**: Measures the accuracy and reliability of the generated answers.
    - **Methodology**: Compare generated answers against a set of ground truth answers for correctness.
2. **Answer Relevance**
    - **Definition**: Evaluates the relevance of the generated answers to the user's query.
    - **Methodology**: Use relevance scoring to assess the match between the generated answers and the query.

3. **Information Integration**
    - **Definition**: Assesses the ability to integrate and present information cohesively.
    - **Methodology**: Evaluate the coherence and completeness of the generated answers.
4. **Counterfactual Robustness**
    - **Definition**: Tests the robustness of the system against counterfactual or contradictory queries.
    - **Methodology**: Introduce counterfactual queries and measure the impact on answer accuracy.
5. **Negative Rejection**
    - **Definition**: Measures the system's ability to reject and handle negative or inappropriate queries.
    - **Methodology**: Introduce negative queries and evaluate the system's response.
6. **Latency**
    - **Definition**: Measures the response time of the system from receiving a query to delivering an answer.
    - **Methodology**: Calculate the time taken for the entire query-response cycle.

3. Methods to Improve Metrics

3.1 Proposed Methods

1. **Improving Context Precision and Recall**
    - **Method**: Enhance the retrieval model using advanced natural language understanding techniques and fine-tuning with a more extensive and diverse dataset.
    - **Implementation**: Implementing a transformer-based retrieval model with fine-tuning.
2. **Enhancing Answer Relevance and Faithfulness**
    - **Method**: Incorporate a post-processing validation step to verify the relevance and faithfulness of the generated answers.
    - **Implementation**: Add a validation layer using a secondary model to cross-check the generated answers.

4. Results and Comparative Analysis

4.1 Baseline Metrics

- **Context Precision**: 0.72
- **Context Recall**: 0.65
- **Context Relevance**: 0.70
- **Context Entity Recall**: 0.68
- **Noise Robustness**: 0.60
- **Faithfulness**: 0.75
- **Answer Relevance**: 0.70
- **Information Integration**: 0.80

- **Counterfactual Robustness**: 0.65
- **Negative Rejection**: 0.70
- **Latency**: 2.5 seconds

4.2 Improved Metrics

- **Context Precision**: 0.80
- **Context Recall**: 0.78
- **Context Relevance**: 0.77
- **Context Entity Recall**: 0.75
- **Noise Robustness**: 0.70
- **Faithfulness**: 0.82
- **Answer Relevance**: 0.78
- **Information Integration**: 0.85
- **Counterfactual Robustness**: 0.72
- **Negative Rejection**: 0.78
- **Latency**: 2.3 seconds

5. Challenges and Solutions

5.1 Challenges

- **Data Diversity**: Ensuring the dataset used for fine-tuning was diverse enough to cover various contexts and queries.
- **Noise Handling**: Effectively introducing and handling noise without significantly degrading performance.

5.2 Solutions

- **Data Augmentation**: Used data augmentation techniques to diversify the training dataset.
- **Advanced Noise Filtering**: Implemented advanced noise filtering techniques to handle irrelevant inputs better.