Report on RAG Pipeline Responses and Performance Evaluation

Project Objectives

The objective of this project is to evaluate the performance of the RAG (Retrieval-Augmented Generation) pipeline and explore methods to enhance its metrics. This includes an analysis of context relevance and answer relevance for a series of questions and generated responses. The performance metrics are analyzed using various metrics such as ROUGE scores, context precision, recall, relevance, and other qualitative measures.

1. Performance Metrics

The RAG pipeline is evaluated based on two main components: context relevance and answer relevance.

1.1. Context Relevance

Context Precision: Measures how accurately the retrieved context matches the user's query.

Context Recall: Evaluates the ability to retrieve all relevant contexts for the user's query.

Context Relevance: Assesses the relevance of the retrieved context to the user's query.

Context Entity Recall: Determines the ability to recall relevant entities within the context.

Noise Robustness: Tests the system's ability to handle noisy or irrelevant inputs.

1.2. Answer Relevance

Accuracy and Relevance: Measures how correct and relevant the responses are to the given prompts.

Coherence and Fluency: Assesses how logical and smoothly flowing the responses are.

Consistency: Evaluates how consistent the responses are over multiple similar queries.

Robustness: Tests how well the model handles ambiguous, tricky, or adversarial inputs.

2. Evaluate Performance

The performance evaluation includes both qualitative and quantitative analysis using human assessment and automated metrics such as cosine similarity for context relevance and ROUGE scores for answer relevance.

2.1. Evaluating Context Relevance

To evaluate context relevance, cosine similarity is used to measure the similarity between the TF-IDF vectors of the questions and the retrieved contexts.

2.2. Evaluating Answer Relevance

Answer relevance is evaluated using ROUGE scores, specifically ROUGE-1 and ROUGE-L, which measure the overlap of unigrams and the longest common subsequence of tokens between the generated and reference answers.

3. Response Analysis and Metrics

Below are the detailed metrics for each query and response pair:

Query 1: What is 'Taxi Blues' about?

Actual Answer: Taxi Blues is a Russian film directed by Pavel Lungin.

Generated Response: The film explores themes of brotherhood, justice, and the human spirit through a series of twists and turns that lead to the uncovering of the truth and ultimately the reconciliation of the three brothers.

Metrics:

- ROUGE-1 F1: 0.0909
- ROUGE-L F1: 0.0455
- Context Precision: 0.0588
- Context Recall: 0.2
- Context Relevance: 0.0
- Context Entity Recall: 0.2
- Noise Robustness: 0
- Faithfulness: 0.0909
- Answer Relevance: 0
- Information Integration: 0.0682
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 3.14 seconds

Query 2: What is 'The Hunger Games' about?

Actual Answer: The Hunger Games is a dystopian science fiction film based on the novel by Suzanne Collins.

Generated Response: The story takes place in a future where North America is divided into multiple warring districts. There are two young people from each district who are selected to take part in the televised events of the games every year and fight until there is only one victor left.

Metrics:

- ROUGE-1 F1: 0.15625

- ROUGE-L F1: 0.09375
- Context Precision: 0.1042
- Context Recall: 0.3125
- Context Relevance: 0.0
- Context Entity Recall: 0.375
- Noise Robustness: 0
- Faithfulness: 0.15625
- Answer Relevance: 0
- Information Integration: 0.125
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 3.18 seconds

Query 3: What is 'Narasimham' about?

Actual Answer: Narasimham is a Malayalam film directed by Shaji Kailas, starring Mammootty.

Generated Response: In a remote village in South India, Narasimham is an old man who is married to four women. When his first wife dies, he brings home another woman to keep everyone happy. But the movie has the genres of Drama, Comedy film

Metrics:

- ROUGE-1 F1: 0.1509
- ROUGE-L F1: 0.1132
- Context Precision: 0.0952
- Context Recall: 0.3636
- Context Relevance: 0.0
- Context Entity Recall: 0.4545
- Noise Robustness: 0
- Faithfulness: 0.1509
- Answer Relevance: 0
- Information Integration: 0.1321
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 3.14 seconds

Query 4: What is 'The Lemon Drop Kid' about?

Actual Answer: The Lemon Drop Kid is a 1951 comedy film starring Bob Hope.

Generated Response: The Lemon Drop Kid is a 1951 American comedy film directed by Frank Tashlin and starring Bob Hope. It is based on the novel The Lemon Drop Kid by Mark Hellinger. Bob plays a New York hustler who learns that if he cannot pay back a debt to the mob, his life will be forfeit. He decides to travel to Hollywood and collect a debt owed him there in order to pay off his mobster creditors.

Metrics:

- ROUGE-1 F1: 0.2727
- ROUGE-L F1: 0.2727
- Context Precision: 0.1579
- Context Recall: 1.0
- Context Relevance: 0.0
- Context Entity Recall: 1.75
- Noise Robustness: 0
- Faithfulness: 0.2727
- Answer Relevance: 0
- Information Integration: 0.2727
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 3.96 seconds

Query 5: What is 'A Cry in the Dark' about?

Actual Answer: A Cry in the Dark is a 1988 Australian film starring Meryl Streep and Sam Neill, based on the true story of Lindy Chamberlain.

Generated Response: (No response generated)

Metrics:

- ROUGE-1 F1: 0.0
- ROUGE-L F1: 0.0
- Context Precision: 0.0
- Context Recall: 0.0
- Context Relevance: 0.0
- Context Entity Recall: 0.0
- Noise Robustness: 0
- Faithfulness: 0.0
- Answer Relevance: 0
- Information Integration: 0.0
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 2.10 seconds

Query 6: What is 'End Game' about?

Actual Answer: End Game is a 2006 South African film directed by Pete Travis.

Generated Response: The film stars Lewis H. Lapham, who plays himself as editor of Harper's Magazine.

Metrics:

- ROUGE-1 F1: 0.0741
- ROUGE-L F1: 0.0741
- Context Precision: 0.0667

- Context Recall: 0.0833
- Context Relevance: 0.0
- Context Entity Recall: 0.0833
- Noise Robustness: 0
- Faithfulness: 0.0741
- Answer Relevance: 0
- Information Integration: 0.0741
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 2.52 seconds

Query 7: What is 'Dark Water' about?

Actual Answer: Dark Water is a 2005 American horror film directed by Walter Salles and based on the Japanese film of the same name.

Generated Response: Dark Water was released in theaters in 2005 and was directed by David Koepp. It's a horror film based on a short story called "Onigiri," which means "cursed rice ball" in Japanese.

Metrics:

- ROUGE-1 F1: 0.4364
- ROUGE-L F1: 0.2909
- Context Precision: 0.3636
- Context Recall: 0.5455
- Context Relevance: 0.0
- Context Entity Recall: 0.5909
- Noise Robustness: 0
- Faithfulness: 0.4364
- Answer Relevance: 0
- Information Integration: 0.3636
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 3.09 seconds

Query 8: What is 'Sing' about?

Actual Answer: Sing is a 2016 animated musical film produced by Illumination Entertainment.

Generated Response: The main character of this film, David (a.k.a. "Kai") was born with a craniofacial condition that required many surgeries throughout his life. He and his family eventually relocated to the United States from China in hopes of finding better treatment for him.

Metrics:

- ROUGE-1 F1: 0.0727

- ROUGE-L F1: 0.0364
- Context Precision: 0.0455
- Context Recall: 0.1818
- Context Relevance: 0.0
- Context Entity Recall: 0.3636
- Noise Robustness: 0
- Faithfulness: 0.0727
- Answer Relevance: 0
- Information Integration: 0.0545
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 3.27 seconds

Query 9: What is 'Meet John Doe' about?

Actual Answer: Meet John Doe is a 1941 American film directed by Frank Capra, starring Gary Cooper and Barbara Stanwyck.

Generated Response: ' Meet John Doe' was a 1946 American film directed by Frank Capra. The movie has the genres of Drama, Suspense, Film noir, Mystery, Gangster Film, Crime Thriller, Thriller, Psychological thriller

Metrics:

- ROUGE-1 F1: 0.4167
- ROUGE-L F1: 0.4167
- Context Precision: 0.3333
- Context Recall: 0.5556
- Context Relevance: 0.0
- Context Entity Recall: 0.5556
- Noise Robustness: 0
- Faithfulness: 0.4167
- Answer Relevance: 0
- Information Integration: 0.4167
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 3.08 seconds

Query 10: What is 'Ghost In The Noonday Sun' about?

Actual Answer: Ghost In The Noonday Sun is a 1973 film starring Peter Sellers.

Generated Response: In an old mine on Christmas Eve, a group of carol singers unwittingly awaken Mr Gruber, a powerful vampire who becomes intent on having them as his slaves. William Rikström (Robert Powell), the only one to escape, is later helped by Gerde (Saskia Wickham) and they decide that someone must be told about the vampires. They eventually track down Father Gregory (Eric Richard), a vampire expert. He tells them of an old prophecy which says that if three people are killed in the name of God then Satan will lose

his power on earth and all vampires, including Mr Gruber, will die. William, Gerde and Father Gregory set out to try to stop Mr Gruber before it is too late.

Metrics:

- ROUGE-1 F1: 0.0602
- ROUGE-L F1: 0.0602
- Context Precision: 0.0331
- Context Recall: 0.3333
- Context Relevance: 0.0
- Context Entity Recall: 0.5
- Noise Robustness: 0
- Faithfulness: 0.0602
- Answer Relevance: 0
- Information Integration: 0.0602
- Counterfactual Robustness: 0
- Negative Rejection: 0
- Latency: 5.30 seconds

4. Conclusion

The evaluation of the RAG pipeline indicates that there is a need for improvement in both context and answer relevance. The system struggles with providing relevant and accurate answers, often diverging significantly from the actual content. The metrics suggest areas for potential enhancement, particularly in context precision, recall, and the overall faithfulness of the responses.

5. Future Work

To improve the RAG pipeline's performance, we will explore the following areas:

- **Enhanced Retrieval Mechanisms**: Improve the retrieval component to ensure more relevant context is provided.
- **Model Training**: Further train the generation model to enhance coherence and relevance of the answers.
- **Noise Handling**: Implement techniques to better handle noisy or irrelevant inputs.
- **Robustness**: Improve the model's robustness against ambiguous or adversarial inputs.

By addressing these areas, we aim to enhance the overall accuracy, relevance, and quality of the RAG pipeline's responses.