

Calculating and Reporting Metrics of the RAG Pipeline in Wedding Music Recommendation Chatbot

Yue Wang
yueyuelala19@gmail.com

July 29, 2024

Contents

1	Performance Metrics Calculation	3
1.1	Retrieval Metrics	3
1.1.1	Assessment	3
1.1.2	Retrieval context	3
1.1.3	Context Precision	4
1.1.4	Context Recall	4
1.1.5	Context Relevance	5
1.1.6	Context Entity Recall	5
1.1.7	Noise Robustness	5
1.2	Generation Metrics	6
1.2.1	Assessment	6
1.2.2	Generation result	6
1.2.3	Faithfulness	6
1.2.4	Answer Relevance	7
1.2.5	Information Integration	7
1.2.6	Counterfactual Robustness	8
1.2.7	Negative Rejection	8
1.3	Latency	8
2	Methods to Improve Metrics	9
2.1	Propose and implement methods to improve metrics	9
2.1.1	Propose	9
2.1.2	Analysis	9
2.1.3	Implement methods	10
2.2	Changes made and analyze their impact on the overall performance of the RAG pipeline.	10
2.2.1	Extend data	10
2.2.2	Improve the QA system	10

2.2.3	Adjust the LLM settings	10
2.3	Impact on the overall performance of the RAG pipeline	10
2.3.1	Retrieval context relevance improved	10
2.3.2	Generation answer's faithfulness, answer relevance, and Information Integration improved	11
3	Negative impact	12
3.1	Latency time	12
4	Challenges faced and addressed	12


```

pineconenotbook — python • streamlit run app.py — 156x24

Test the version1 context=====
=====chatbot v1 retrieval context and generated answer=====
Query: What's the most popular song 2023?

Generated Answer: I don't have information on the most popular song in 2023.

Retrieve Contexts [Document(metadata={'Header 2': 'Happy'}, page_content='## Happy\n- **Artist**: Pharrell Williams\n- **Genre**: Pop\n- **Year**: 2013\n- **Description**: An upbeat and joyful song perfect for the reception.\n- **Popularity**: 90\n- **Duration (ms)**: 233000\n- **Explicit**: No\n- **Album**: G I R L'), Document(metadata={'Header 2': 'Happy'}, page_content='## Happy\n- **Artist**: Pharrell Williams\n- **Genre**: Pop\n- **Year**: 2013\n- **Description**: An upbeat and joyful song perfect for the reception.\n- **Popularity**: 90\n- **Duration (ms)**: 233000\n- **Explicit**: No\n- **Album**: G I R L'), Document(metadata={'Header 2': 'Happy'}, page_content='## Happy\n- **Artist**: Pharrell Williams\n- **Genre**: Pop\n- **Year**: 2013\n- **Description**: An upbeat and joyful song perfect for the reception.\n- **Popularity**: 90\n- **Duration (ms)**: 233000\n- **Explicit**: No\n- **Album**: G I R L'), Document(metadata={'Header 2': 'Happy'}, page_content='## Happy\n- **Artist**: Pharrell Williams\n- **Genre**: Pop\n- **Year**: 2013\n- **Description**: An upbeat and joyful song perfect for the reception.\n- **Popularity**: 90\n- **Duration (ms)**: 233000\n- **Explicit**: No\n- **Album**: G I R L')]

```

1.1.3 Context Precision

- Definition: The ratio of relevant context retrieved to the total context retrieved.
- Evaluation: For queries like “Can you recommend some songs for a beach wedding ceremony?” and “What are some classical music suitable for the first dance?,” the retrieved context should match closely with wedding music recommendations. If the chatbot retrieves specific song details relevant to beach weddings or classical music from the dataset, the precision is high.
- Assessment:
 - Query 1: low
 - Query 2: low
 - Query 3: low

1.1.4 Context Recall

- Definition: The ratio of relevant context retrieved to the total relevant context available.
- Evaluation: The chatbot should retrieve all relevant entries about beach weddings and classical music from the dataset.
- Assessment:
 - Query 1: If all relevant beach wedding music entries are retrieved, recall is high. – low
 - Query 2: If all relevant classical music entries are retrieved, recall is high. – low
 - Query 3: Not applicable due to the context being out of scope.

1.1.5 Context Relevance

- Definition: The relevance of the retrieved context to the user’s query.
- Evaluation: The chatbot should provide highly relevant contexts that directly answer the user’s query.
- Assessment:
 - Query 1: Retrieved contexts are relevant to beach wedding recommendations. – low
 - Query 2: Retrieved contexts are relevant to classical music recommendations. – low
 - Query 3: Context relevance is expected to be low as it falls outside the dataset. – low

1.1.6 Context Entity Recall

- Definition: The ability to recall relevant entities (e.g., song titles, artists) within the context.
- Evaluation: The chatbot should recall and present song titles, artists, and other relevant details accurately.
- Assessment:
 - Query 1: High entity recall if specific song titles and artists for beach weddings are presented. – low
 - Query 2: High entity recall if specific classical music titles and artists are presented. – high
 - Query 3: Not applicable.

1.1.7 Noise Robustness

- Definition: The system’s ability to handle noisy or irrelevant inputs.
- Evaluation: The chatbot should maintain accuracy even if the input query contains irrelevant or noisy information.
- Assessment:
 - Query 1: High noise robustness if the system can still provide accurate recommendations despite minor irrelevant details. – low
 - Query 2: High noise robustness under similar conditions. – low
 - Query 3: Should gracefully handle irrelevant queries and respond appropriately. – low

1.2 Generation Metrics

1.2.1 Assessment

- Query 1: “Can you recommend some songs for a beach wedding ceremony?”
- Query 2: “What are some classical music suitable for the first dance?”
- Query 3: “What’s the most popular song 2023?”

1.2.2 Generation result

- Query 1:



I don't have access to the specific songs in the dataset to recommend for a beach wedding ceremony.



Can you recommend some songs for a beach wedding ceremony?

- Query 2:



For a classical music selection suitable for the first dance, you might consider "The Way You Look Tonight" by Frank Sinatra. It's a classic love song perfect for such a moment.



What are some classical music suitable for the first dance?

- Query 3:



I don't have information on the most popular song in 2023.



What’s the most popular song 2023?

1.2.3 Faithfulness

- Definition: The accuracy and reliability of the generated answers.

- Evaluation: The generated answers should be accurate and based on the provided dataset.
- Assessment:
 - Query 1: High faithfulness if the recommendations match the dataset entries. – low
 - Query 2: High faithfulness if the classical music recommendations match the dataset. – low
 - Query 3: The response should correctly identify the lack of information in the dataset. – high

1.2.4 Answer Relevance

- Definition: The relevance of the generated answers to the user’s query.
- Evaluation: The answers should be directly relevant to the user’s queries about wedding music.
- Assessment:
 - Query 1: High relevance with specific beach wedding music recommendations. – low
 - Query 2: High relevance with specific classical music recommendations. – low
 - Query 3: Response should explain the lack of relevant data. – high

1.2.5 Information Integration

- Definition: The ability to integrate and present information cohesively.
- Evaluation: The chatbot should present information in a coherent and integrated manner.
- Assessment:
 - Query 1: High if the response integrates multiple relevant songs coherently. – low
 - Query 2: High if the response integrates relevant classical music suggestions. – low
 - Query 3: Not applicable.

1.2.6 Counterfactual Robustness

- Definition: The robustness of the system against counterfactual or contradictory queries.
- Evaluation: The chatbot should handle contradictory queries gracefully without providing incorrect information.
- Assessment:
 - Query 1: High robustness if it avoids contradictions in the recommendation. – low
 - Query 2: High robustness under similar conditions. – low
 - Query 3: Should clearly identify the lack of relevant data. – high

1.2.7 Negative Rejection

- Definition: The system's ability to reject and handle negative or inappropriate queries.
- Evaluation: The chatbot should reject inappropriate queries politely and professionally.
- Assessment:
 - Query 1: Not applicable.
 - Query 2: Not applicable.
 - Query 3: High if the response politely redirects the user to relevant queries. – high

1.3 Latency

- Definition: The response time of the system from receiving a query to delivering an answer.
- Evaluation: The system should have low latency, providing responses quickly.
- Assessment: Measure the response time for each query.
 - Average latency: 2.76
 - Query 1: response time: 1.84

Generated Answer:
{'query': 'Can you recommend some songs for a beach wedding ceremony?', 'result': 'I don't have access to the specific songs in the dataset to recommend for a beach wedding ceremony.', 'source_documents': [Document(metadata={'Header 2': 'Wedding Music Recommendation Dataset'}, page_content='## Wedding Music Recommendation Dataset \nThis dataset is designed to provide a selection of songs suitable for weddings, encompassing various genres and moments within the event, such as the ceremony, reception, and first dance.'), Document(metadata={'Header 2': 'Wedding Music Recommendation Dataset'}, page_content='## Wedding Music Recommendation Dataset \nThis dataset is designed to provide a selection of songs suitable for weddings, encompassing various genres and moments within the event, such as the ceremony, reception, and first dance. \n### Music Entries'), Document(metadata={'Header 2': 'Wedding Music Recommendation Dataset'}, page_content='## Wedding Music Recommendation Dataset \nThis dataset is designed to provide a selection of songs suitable for weddings, encompassing various genres and moments within the event, such as the ceremony, reception, and first dance. \n### Music Entries'), Document(metadata={'Header 2': 'Wedding Music Recommendation Dataset'}, page_content='## Wedding Music Recommendation Dataset \nThis dataset is designed to provide a selection of songs suitable for weddings, encompassing various genres and moments within the event, such as the ceremony, reception, and first dance. \n### Music Entries')]}
Response Time: 1.841718262298584 seconds

– Query 2: response time 4.19

```
Generated Answer:
{'query': 'What are some classical music suitable for the first dance', 'result': 'If you are looking for classical music suitable for the first dance at a wedding, you may want to consider pieces like "Canon in D" by Johann Pachelbel, "Clair de Lune" by Claude Debussy, or "Meditation from Thaïs" by Jules Masse net. Classical pieces like these are often popular choices for romantic and elegant first dances.', 'source_documents': [Document(metadata={'Header 2': 'Wedding Music Recommendation Dataset'}, page_content='## Wedding Music Recommendation Dataset \nThis dataset is designed to provide a selection of songs suitable for weddings, encompassing various genres and moments within the event, such as the ceremony, reception, and first dance.'), Document(metadata={'Header 2': 'The Way You Look Tonight'}, page_content='## The Way You Look Tonight\n\n **Artist**: Frank Sinatra\n\n **Genre**: Jazz\n\n **Year**: 1964\n\n **Description**: A classic love song perfect for the first dance.\n\n **Popularity**: 90\n\n **Duration (ms)**: 210000\n\n **Explicit**: No\n\n **Album**: Sinatra Sings Days of Wine and Roses, Moon River, and Other Academy Award Winners'), Document(metadata={'Header 2': 'The Way You Look Tonight'}, page_content='## The Way You Look Tonight\n\n **Artist**: Frank Sinatra\n\n **Genre**: Jazz\n\n **Year**: 1964\n\n **Description**: A classic love song perfect for the first dance.\n\n **Popularity**: 90\n\n **Duration (ms)**: 210000\n\n **Explicit**: No\n\n **Album**: Sinatra Sings Days of Wine and Roses, Moon River, and Other Academy Award Winners'), Document(metadata={'Header 2': 'The Way You Look Tonight'}, page_content='## The Way You Look Tonight\n\n **Artist**: Frank Sinatra\n\n **Genre**: Jazz\n\n **Year**: 1964\n\n **Description**: A classic love song perfect for the first dance.\n\n **Popularity**: 90\n\n **Duration (ms)**: 210000\n\n **Explicit**: No\n\n **Album**: Sinatra Sings Days of Wine and Roses, Moon River, and Other Academy Award Winners')]]
Response Time: 4.197391626205444 seconds
```

– Query 3: response time 2.25

```
Generated Answer:
{'query': 'What's the most popular song 2023?', 'result': 'I don't have information on the most popular song in 2023.', 'source_documents': [Document(metadata={'Header 2': 'Happy'}, page_content='## Happy\n\n **Artist**: Pharrell Williams\n\n **Genre**: Pop\n\n **Year**: 2013\n\n **Description**: An upbeat and joyful song perfect for the reception.\n\n **Popularity**: 90\n\n **Duration (ms)**: 233000\n\n **Explicit**: No\n\n **Album**: G I R L'), Document(metadata={'Header 2': 'Happy'}, page_content='## Happy\n\n **Artist**: Pharrell Williams\n\n **Genre**: Pop\n\n **Year**: 2013\n\n **Description**: An upbeat and joyful song perfect for the reception.\n\n **Popularity**: 90\n\n **Duration (ms)**: 233000\n\n **Explicit**: No\n\n **Album**: G I R L'), Document(metadata={'Header 2': 'Happy'}, page_content='## Happy\n\n **Artist**: Pharrell Williams\n\n **Genre**: Pop\n\n **Year**: 2013\n\n **Description**: An upbeat and joyful song perfect for the reception.\n\n **Popularity**: 90\n\n **Duration (ms)**: 233000\n\n **Explicit**: No\n\n **Album**: G I R L'), Document(metadata={'Header 2': 'Happy'}, page_content='## Happy\n\n **Artist**: Pharrell Williams\n\n **Genre**: Pop\n\n **Year**: 2013\n\n **Description**: An upbeat and joyful song perfect for the reception.\n\n **Popularity**: 90\n\n **Duration (ms)**: 233000\n\n **Explicit**: No\n\n **Album**: G I R L')]]
Response Time: 2.2505762577056885 seconds
```

2 Methods to Improve Metrics

2.1 Propose and implement methods to improve metrics

2.1.1 Propose

Improve the retrieval context relevance, generation answer's faithfulness, answer relevance, and Information Integration.

2.1.2 Analysis

It is normal in some cases for the generated result to recommend something that wasn't explicitly retrieved in the top results, especially if the retrieval step did not find highly relevant contexts. However, in a well-functioning Retrieval-Augmented Generation (RAG) system, the retrieved contexts should ideally inform the generated answer closely. Understanding the discrepancy:

- Retrieval:
 - The retrieval step fetches documents or contexts that are semantically close to the query. However, in your example, it seems that the retrieval contexts are about a jazz song by Frank Sinatra rather than classical music, which indicates a potential mismatch in the retrieval process.
- Generation:
 - The generation step uses the retrieved contexts to craft a response. If the retrieved contexts do not adequately cover the query's topic, the language model might generate an answer based on its training data, which includes general knowledge, rather than the specific retrieved contexts.

2.1.3 Implement methods

To address the retrieval performance, we need to ensure that the retrieval step is more effective in fetching relevant contexts.

- Ensure Diverse and Relevant Data:
 - Verify that the data indexed in Pinecone includes a diverse range of entries covering different types of music suitable for various wedding moments.
- Improve the QA system:
 - The RetrievalQA class is deprecated, use the newest retrieve chain can help improve the relevance of the retrieved contexts.
- Increase the Number of Retrieved Contexts:
 - Retrieve more contexts (e.g., top_k=3 or more) and filter them for relevance before passing them to the generation step.

To address the generation performance, we can try to adjust the LLM related arguments, such as adjust the model type and temperature value.

2.2 Changes made and analyze their impact on the overall performance of the RAG pipeline.

2.2.1 Extend data

Add more diverse range of entries covering different types of music, and keep the dataset all entries, remove the introduction and conclusion to reduce distractions during retrieve.

2.2.2 Improve the QA system

Integrate the create_retrieval_chain, create_stuff_documents_chain, ChatPromptTemplate, and adjust top_k to 3 to improve the QA system.

2.2.3 Adjust the LLM settings

Adjust model from gpt-3.5-turbo to the newest gpt-4o-mini, change the temperature from 0.2 to 0.4 to make the LLM more creative.

2.3 Impact on the overall performance of the RAG pipeline

2.3.1 Retrieval context relevance improved

The retrieve context of query “Can you recommend some songs for a beach wedding ceremony” is “Can’t Help Falling in Love”, instead of the wedding music dataset introduction before.

```
Calculating-and-Reporting-Metrics-of-the-RAG-Pipeline — python • streamlit run app.py — 155x35
Query: Can you recommend some songs for a beach wedding ceremony?

Generated Answer: Absolutely! A beach wedding ceremony often calls for music that complements the serene and romantic atmosphere. Here are a few song recommendations that would be perfect for such an occasion:

1. "Can't Help Falling in Love" by Elvis Presley - This classic love song is timeless and evokes strong emotions, making it a beautiful choice for your ceremony.
2. "Somewhere Over the Rainbow" by Israel Kamakawiwo'ole - This gentle ukulele rendition adds a dreamy vibe that fits perfectly with a beach setting.
3. "Marry Me" by Train - A sweet and heartfelt song that captures the essence of love and commitment.
4. "Better Together" by Jack Johnson - With its laid-back acoustic style, this song is ideal for a relaxed beach atmosphere.
5. "A Thousand Years" by Christina Perri - This romantic ballad has become a popular choice for weddings, expressing everlasting love.

These songs can help create a magical ambiance for your beach wedding ceremony. If you need more suggestions or specific details about any of these songs, feel free to ask!

Retrieve Contexts: [Document(metadata={'Header 2': 'Can't Help Falling in Love'}, page_content='## Can't Help Falling in Love\n\n- **Artist:** Elvis Presley\n- **Genre:** Rock\n- **Year:** 1961\n- **Description:** A classic love song perfect for any wedding moment.\n- **Popularity:** 94\n- **Duration (ms):** 178000\n- **Explicit:** No\n- **Album:** Blue Hawaii'), Document(metadata={'Header 2': 'Can't Help Falling in Love'}, page_content='## Can't Help Falling in Love\n\n- **Artist:** Elvis Presley\n- **Genre:** Rock\n- **Year:** 1961\n- **Description:** A classic love song perfect for any wedding moment.\n- **Popularity:** 94\n- **Duration (ms):** 178000\n- **Explicit:** No\n- **Album:** Blue Hawaii'), Document(metadata={'Header 2': 'Can't Help Falling in Love'}, page_content='## Can't Help Falling in Love\n\n- **Artist:** Elvis Presley\n- **Genre:** Rock\n- **Year:** 1961\n- **Description:** A classic love song perfect for any wedding moment.\n- **Popularity:** 94\n- **Duration (ms):** 178000\n- **Explicit:** No\n- **Album:** Blue Hawaii')]
```

2.3.2 Generation answer’s faithfulness, answer relevance, and Information Integration improved

When send the query “What are some classical music suitable for the first dance?”, chatbot will give the music in genre classical, instead of jazz before, but also give a jazz as an alternative choice, but will mention it’s genre is jazz. And integrate 3 pieces of classical music to user.

Chatbot_v2

Welcome! How can I assist you today?

Enter your question:

What are some classical music suitable for the first dance?

Press Enter to submit form

Ask

While my expertise is primarily in wedding music planning, I can suggest some beautiful classical pieces that are often chosen for first dances. Here are a few timeless options:

- 1. **"Clair de Lune" by Claude Debussy** - This piece is soft and romantic, creating a dreamy atmosphere.
- 2. **"The Swan" from "Carnival of the Animals" by Camille Saint-Saëns** - A graceful and elegant choice that captures the essence of love.
- 😊 3. **"Ave Maria" by Franz Schubert** - A serene and heartfelt piece that can add a touch of spirituality to your first dance.

If you're looking for something more contemporary, I recommend considering "The Way You Look Tonight" by Frank Sinatra. It's a classic love song that's perfect for a first dance, blending jazz with romantic vibes.

3 Negative impact

3.1 Latency time

- Query 1: 9.52
- Query 2: 7.78
- Query 3: 3.17
- Average latency time: 6.82
- Significant increase in latency, almost 2.5 times that of version 1

4 Challenges faced and addressed

- No ideas on how to measure individual indicators, especially since the evaluation of many of them seems to be very subjective.
 - Enhance my understanding by googling and GPT, and referring to the blog posts provided by the TA.
- Want to print the text retrieved for each question, but don't know how to get.
 - Look up the official documentation of LangChain and Pinecone to find the corresponding class and usage examples and try.
- When getting the first version of the retrieved text, it prompts an error about exceeding the length of the token. It turns out that in the first version I used the prompt and the user input together as a query to do the retrieval, which resulted in a long query to retrieve and caused the error.
 - Adjust only send the user input to retrieve relevant splits.