Chatbot Evaluation Report

1. Overview

This report evaluates the performance of a chatbot based on its responses to user queries. The chatbot's performance is assessed using three primary evaluation metrics: **Accuracy**, **Relevance**, and **Satisfaction**. Additionally, more advanced metrics such as **BLEU**, **ROUGE**, and **METEOR** scores are calculated to assess the quality of the chatbot's responses in comparison to predefined answers from a CSV file.

2. Evaluation Methodology

The chatbot responses were compared to predefined answers that were manually crafted to match the expected responses for a set of common questions. The evaluation was carried out using the following metrics:

- **Accuracy**: Measured how closely the chatbot's response matched the predefined answer.
- **Relevance**: Determined if the chatbot's response was related and relevant to the user's query.
- **Satisfaction**: Evaluated based on the clarity and length of the response (longer answers are assumed to provide more clarity).

Additionally, three NLP evaluation scores were computed:

- **BLEU (Bilingual Evaluation Understudy Score)**: Measures the overlap between the bot's response and the reference answer.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: Measures the overlap of n-grams, which reflects how similar the content of the bot's response is to the reference.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)**: A metric to evaluate the quality of the response by measuring precision, recall, synonymy, stemming, and word order.

3. Evaluation Results

User Question 1: What is the return policy for items purchased at our store?

- **Bot Response**: Our store offers a comprehensive return policy designed to make your shopping experience hassle-free. (includes detailed policy breakdown)
- **Predefined Answer**: You can return most items within 30 days of purchase for a full refund or exchange. Items must be in their original condition, with all tags and packaging intact. Please bring your receipt or proof of purchase when returning items.
    - **Accuracy**: 1.7 / 5
    - **Relevance**: 3 / 5

    - **Satisfaction**: 5 / 5
    - **BLEU Score**: 19.20
    - **ROUGE Scores**:
        - rouge1: 0.4242
        - rouge2: 0.3558
        - rougeL: 0.4
    - **METEOR Score**: None (Error due to tokenization issue)

User Question 2: Are there any items that cannot be returned under this policy?

- **Bot Response**: Yes, there are certain items that are non-returnable under our return policy. (details specific non-returnable items)
- **Predefined Answer**: Yes, certain items such as clearance merchandise, perishable goods, and personal care items are non-returnable. Please check the product description or ask a store associate for more details.
    - **Accuracy**: 0.1 / 5
    - **Relevance**: 3 / 5
    - **Satisfaction**: 5 / 5
    - **BLEU Score**: 16.82
    - **ROUGE Scores**:
        - rouge1: 0.3619
        - rouge2: 0.2654
        - rougeL: 0.3478
    - **METEOR Score**: None (Error due to tokenization issue)

User Question 3: How will I receive my refund?

- **Bot Response**: Refunds will be issued to the original form of payment used at the time of purchase. (details refund process)
- **Predefined Answer**: Refunds will be issued to the original form of payment used at the time of purchase. Please bring your receipt or proof of purchase for processing.
    - **Accuracy**: 1.21 / 5
    - **Relevance**: 3 / 5
    - **Satisfaction**: 5 / 5
    - **BLEU Score**: 20.15
    - **ROUGE Scores**:
        - rouge1: 0.4347
        - rouge2: 0.3685
        - rougeL: 0.4123
    - **METEOR Score**: None (Error due to tokenization issue)

4. Key Observations

- **Accuracy**: The chatbot's accuracy in matching predefined answers was generally low (ranging from 0.1 to 1.7 out of 5). This indicates that while the chatbot provided helpful information, it did not fully align with the exact phrasing of the predefined answers.
- **Relevance**: The relevance of the responses was rated at 3 out of 5 for most cases. This suggests that the chatbot's answers were somewhat related but might have included excessive or unnecessary details in certain situations.
- **Satisfaction**: The satisfaction score remained high (5/5), as the chatbot's responses were deemed clear and comprehensive. However, this could be due to the length and structure of the answers, which were detailed.
- **BLEU Score**: The BLEU scores ranged from 16 to 20, indicating moderate overlap between the chatbot's responses and the predefined answers. While not perfect, the responses were somewhat aligned in terms of content.
- **ROUGE Scores**: The ROUGE scores indicate a fair level of similarity in terms of word overlap (rouge1 and rougeL scores) but indicate that the responses were not exact matches.
- **METEOR Score**: There were errors in calculating the METEOR score, mainly due to tokenization issues with the input sentences. This highlights a need for preprocessing the responses for evaluation.

5. Conclusion and Recommendations

- The chatbot's responses were generally helpful, but they lacked a high degree of precision when compared to predefined answers. There is room for improvement in terms of accuracy and alignment with expected responses.
- While the chatbot's responses were relevant and clear, they could be refined to better match the expected language used in predefined answers.
- Future iterations of the chatbot should include improvements in tokenization and response generation, which may help to better handle metrics like METEOR.
- A better handling of variations in language and response phrasing could improve both the accuracy and relevance scores.

Overall, the chatbot performed reasonably well in providing relevant information and the response is good this chatbot is an intelligence bot and it is not expected to give the same or exact answer to the question that is why i cannot judge it base the accuracy, relevance but base on the satisfaction, BLEU Score and ROUGE Scores it’s perform very well