Conversational Agent

Eviatar Nachshoni, M.Sc Computer Science, BIU, ONLP Lab

July 2024

1 Evaluation

1.1 Building Test Set

To create the test set, we used Mixtral-8x22B-Instruct-v0.1 to generate 10 dialogues, each with a varying number of turns. We then reviewed and filtered these dialogues to eliminate any errors and ensure high quality.

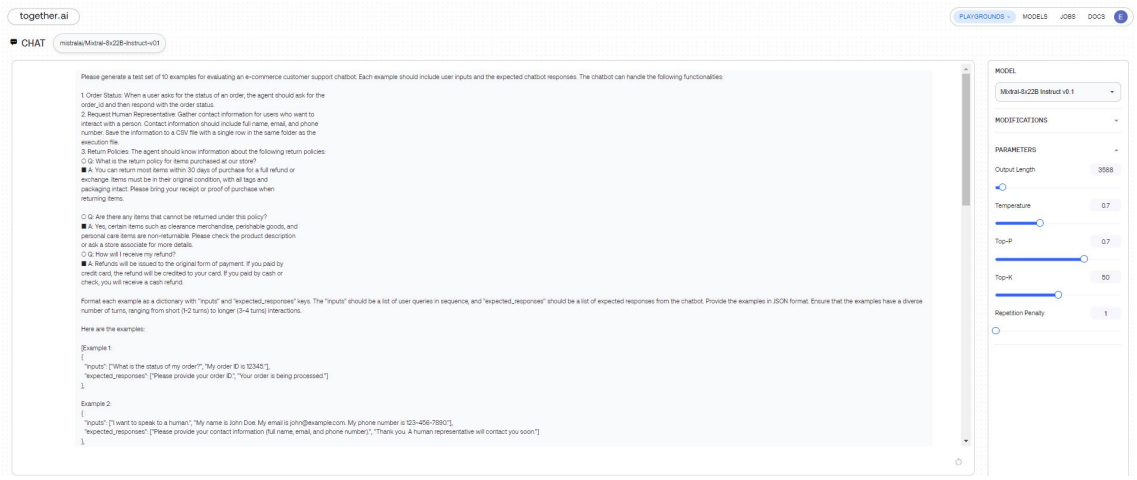

Figure 1: The prompt provided to Mistral22*8b.

1.2 Evaluation Methods

The methods we used to evaluate the chatbot's performance are outlined below:

- **Accuracy**
  - **Simulation of Conversations:**
    - Each predefined test case from the JSON file is run through the `react_agent_chat` function to obtain the chatbot's actual responses.

- **BERTScore Evaluation:**
  - Each actual response from the chatbot is compared to the expected response using BERTScore.
  - BERTScore evaluates the similarity between the two texts at a semantic level, providing precision, recall, and F1 scores.
- **Determining Correctness:**
  - A response is considered correct if its BERTScore F1 score is above a specified threshold (default is 0.85).
- **Accuracy Calculation:**
  - The total number of correct responses is divided by the total number of responses to get the accuracy.
  - The formula used is:

$$\text{Accuracy} = \left(\frac{\text{Number of Correct Responses}}{\text{Total Number of Responses}} \right) \times 100\%$$

- **Response Relevance**
  - I manually annotated and compared the model's expected output to the model's actual output to check if the answer is relevant to the question. The annotations are in the `evaluation_results.xlsx` file.
- **User Satisfaction**
  - I manually annotated and compared the model's expected output to the model's actual output to check if it provided enough information. The annotations are in the `evaluation_results.xlsx` file.

1.3 Evaluation Results

Metric	**Mixtral-8x22B-Instruct-v0.1**
Accuracy	96% [1]
Relevance	96%
User Satisfaction	92%

Table 1: Evaluation results for the model based on different evaluation methods.