Introduction

This evaluation report outlines the development and testing of a conversational agent designed to handle customer support queries for an e-commerce platform. The primary objective of this project was to create a chatbot capable of managing multi-turn conversations and delivering accurate responses to common customer inquiries, such as order status, return policies, and other support-related topics.

The chatbot was implemented using the OpenAI API, leveraging the capabilities of a large language model (LLM) to provide a robust and intelligent conversational experience. Its performance was assessed using a set of predefined dialogues to ensure effectiveness in addressing typical customer needs.

This report provides an overview of the implementation process, testing outcomes, and insights gathered during the evaluation.

Functionalities

1. **Order Status**
    - The agent requests the `order_id` from the user and provides the corresponding order status.
2. **Request Human Representative**
    - The agent collects contact information (full name, email, and phone number) from users who wish to interact with a human representative.
    - The information is saved in a CSV file named after the user in the same folder as the execution file.
3. **Return Policies**
    - The agent provides detailed information about return policies

EVALUATION

The evaluation of the conversational agent will be conducted based on three key criteria:

1. **Accuracy**: The ability of the chatbot to provide correct responses to user queries.
2. **Response Relevance**: Ensuring that responses are contextually appropriate and address the user's specific concerns.
3. **Bot Adaptability**: Assessing the chatbot’s capacity to handle diverse user inputs, including variations in tone, phrasing, and intent, while satisfying user needs effectively.

Test Scenarios

The evaluation process will utilize a set of predefined conversations, divided into **basic** and **advanced** categories, to assess the chatbot's performance comprehensively.

Basic Conversations

The basic scenarios involve straightforward questions and typical user interactions:

- "Hi there, what is the status for order SAV_5?"
- "Hi there, what is the status for order SAV_190?"
- "What is the return policy for items purchased at our store?"
- "Are there any items that cannot be returned under this policy?"
- "How will I receive my refund?"
- "I want to speak to a rep."
- "My name is Christopher, chris@gmail.com, 08154481095."
- "Hi, my order is getting delayed. I want to talk to someone about this."
- "My name is Kendrick Lamar, kdot@gmail.com, 08154481098."

Advanced Conversations

The advanced scenarios involve more nuanced or challenging interactions, designed to test the chatbot's adaptability to indirect, vague, or emotionally charged inputs:

- "Hi, is my package status for order SAV_21 ever going to be completed or are the workers asleep?" (rude tone)
- "I purchased some watches last week. Can I still return them?" (less direct)
- "How about apples?" (deviating from return policy context)
- "I paid by card but can I get a cash refund? Also, I don't feel comfortable talking to a non-human." (deviating from policy and indirect rep request)
- "My name is Jack Harlow, jack@gmail.com, 08154481095." (follow-up for context)

This structured evaluation ensures that the chatbot is rigorously tested for both basic and complex interactions, providing insights into its practical effectiveness and user satisfaction.

Accuracy

In regards to generating order status, The agent always showed accuracy in calling the necessary tools in the 3 cases for checking order status, the model had no problem understanding the calling the correct tools and generating an accurate overall response

In the 6 conversations regarding policies, the agent perfectly references the Q and A while incorporating necessary information from the use

In the conversations requesting for human representatives, the agent demonstrated recall when incomplete information was passed, asked for clarification before calling the necessary tools

Response Relevance and User Satisfaction from interaction

In the more advanced conversations, the agent was able to navigate through emotions remarks clouding the requests and still address the core issue as seen in the video demo

Also, the agent had no problem understanding chained requests and in parallel addressing both requests as seen in the penultimate and last conversation in the advanced conversation set

Conclusion

In addressing the needs of handling customer support queries, a conversational AI agent was created in fulfilling 3 major request with a high degree of accuracy:

Order Status

Request Human Representative

Return Policies

All 3 functionalities were then successfully tested and evaluated with varying degree of difficulty.