Task: Customer Support Chatbot  
Use Cases: Customer support for common issues customers face when using or navigating websites

Objectives: Generate satisfactory responses to customers’ questions

Prompt Flow Design Components:

Input 1 (question): A customer support question

Input 2 (context): Provides example customer support questions

FAQ (LLM): Generates a prompt with the given inputs, passes it onto the model, and generates a text response

Echo (Python): Outputs the response generated

Output (response): prints the answer generated by the model and retrieved through echo.

Prototype Summary:

Question: When is the latest date I can return a purchased item?

Response: “The latest date you can return a purchased item is 30 days from the date of purchase, provided the item is in its original condition. Please note that some items, such as final sale items, may not be eligible for return. If you have any further questions or need assistance with your return, feel free to contact customer service.”

Problems:

FAQ only has one input, but I want to use two inputs. Once I ran the promptflow, the inputs section below list “question” and “context” since I removed the “topic” input and added the former two on the code.

Because I renamed “joke” LLM to FAQ, echo did not work properly since it expected a response from “joke”. I solved this by switching the expected input to FAQ.output

I needed to make a custom dataset with two columns for evaluation. I was able to do this by making an excel spreadsheet with the data and saving it as a csv.

Evaluation:

Evaluations used: QnA F1 score evaluation, QnA Groundedness, QnA Relevance, QnA Fluency

F1 score (5-question context): 0.36

Groundedness (5-question context): Failed due to line execution timeout

QnA Relevance (5-question context): Failed due to line execution timeout

QnA Fluency (5-question context): 5.00

QnA Coherence (5-question context): 4.5

QnA GPT Similarity: Failed due to line execution timeout

Evaluation analysis: Model seems to be performing good, but it could use some noticeable improvements. Although I deployed the promptflow and configured a monitor for the endpoint, the monitor does not display any data. The logs do indicate that the endpoint is receiving input, processing the input, and returning a response. This process takes three seconds. The endpoint has not experienced any errors so far. When I gave a prompt that told the model to tell a joke, the response stated that the endpoint was for FAQs, but tried to tell a joke anyway

Future Improvements:

Editing the context to give the response a one-paragraph limit

Using more questions for the fine tuning

Increasing the number of questions for the context

Allocating more powerful compute cores to the promptflow

Reject any irrelevant or non-questions.