**Human Evaluation Strategy**

The human evaluation study for our GrocerEase chatbot will involve creating a structured approach to gather qualitative and quantitative feedback from users.

**Study I) Comparing GrocerEase Chatbot with Baseline (Amazon or Instacart) *In-Person***

To evaluate and compare the efficiency, user satisfaction, and accuracy of grocery shopping tasks between the GrocerEase chatbot interface and the search engine interfaces of Amazon and Whole Foods.

### **Hypotheses (We can define objectives/pillars eg. accuracy/relevance of response, query understanding, user satisfaction, overall satisfaction)**

1. The GrocerEase chatbot will provide a more efficient shopping experience than traditional search query based ecomm/online grocery platforms
2. Users will report higher satisfaction with the GrocerEase chatbot due to its conversational interface. (Q1)
3. The GrocerEase chatbot will demonstrate higher accuracy in understanding and responding to specific grocery queries. (Q2)

### **Task Design (To simulate real life needs while online shopping)**

1. **Find Specific Items for occasion**: For example: “Make a list of things to carry for a 3 day hike (e.g., organic apples, gluten-free bread)”. (Q: The users will ask for a basket containing 3 items and will rate the basket they prefer - binary decision variable).
2. **Budget-Based Shopping**: Create a shopping list for a meal under $50.
3. **Dietary Specific Shopping**: Find items that meet a particular dietary requirement (e.g., vegan, keto or low calorie (<200) (Q2)
4. **Compare Products**: Compare two similar products based on price, brand, and nutritional value.
5. **New Item Discovery:** The user asks for 3 items aligned with an occasion and captures the number of items they have never bought/used before. (Q3 -edit)

### **The Experiment (In person)**

### **Brief participants** about the study and give a brief demo of how to use GrocerEase (and obviously amazon/wholefoods if they do not know)

### **Retrieve a list of items for various tasks on both platforms**

### Execute task on amazon/wholefoods, participants are free to use any search query on those website multiple times and are free to get a list of those items however they need

### Now execute same task and prepare a list for specified scenarios using GrocerEase

* **Time Tracking:** Record the time taken to complete each task on both system

### **Post Experiment Brief Survey**

Post-task surveys with Likert scale questions and open-ended responses to gauge user satisfaction, perceived ease of use, task difficulty , time taken etc. We can do data analysis of this to confirm our hypotheses

**(AND/OR)**

**Study II) Questionnaire (both Likert Scale & open-ended questions) - *Offline***

1. **Objectives (**in progress, can add more)

To evaluate the GrocerEase chatbot in terms of its :

* Accuracy of Responses (e.g. were relevant grocery items retrieved?)
* Understanding User Queries (e.g. did you think grocerease understood your request?)
* Overall satisfaction
* etc.

1. **Scenarios (**in progress, can add more)

**Hike Preparation**: Finding suitable snack options and portable foods for a hiking trip.

**Halloween Party**: Organizing a Halloween party shopping list, including decorations and themed foods.

**Thanksgiving Dinner**: Planning a Thanksgiving dinner grocery list, considering dietary preferences and budget.

**Recipe Assistance**: Selecting ingredients for a specific recipe, with substitutions for allergies or preferences.

**Product Recommendations**: Seeking recommendations for specific products like organic, gluten-free, or dairy-free items.

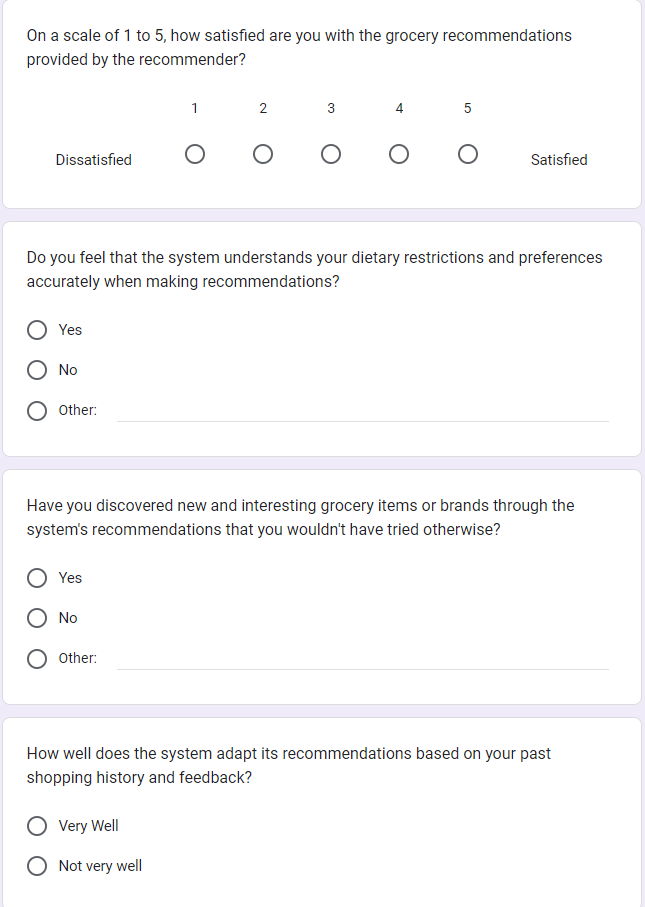
**General Grocery List**: Creating a weekly grocery shopping list based on family size and dietary needs.

1. **Google Form Link**

* **Introduction:** Brief participants on the purpose of the study and how to interact with the GrocerEase chatbot.
* **Scenario Execution:** Participants will interact with the chatbot, completing tasks based on the provided scenarios.
* **Questionnaire:** Subsequently, users will answer questions
* Likert Scale Questions: To quantitatively assess aspects like ease of use, accuracy of responses, and understanding of queries (e.g. how relevant were the items retrieved?, how easy was it?)
* Open-Ended Questions: To gather qualitative feedback on user experience, likes, dislikes, and suggestions for improvement (e.g. What did you like most about the chatbot?What improvements would you suggest? Any additional comments or feedback)

1. **Data Analysis**

Quantitative data from Likert-scale questions can be statistically analyzed to identify performance, satisfaction, patterns, strengths, and areas for improvement, while qualitative data from open-ended questions can be reviewed to extract common themes and insights.



Todo: Answer this questions for AWS, then compare.