# LLMs Project Group Running Meeting Notes GrocerEase

## 12/3

Action Items:

* Everyone: try out 10 one-time prompts each that fit into our main use cases by tuesday mid-day
* Ana will send us questions for human eval by noon tomorrow
* Everyone: poster sections by tuesday night

## 11/30

Action Items:

* Perform auto evaluation - Noel
* Create dataset - Ameya
* Beautify front end - Ameya
* Ana - Run the human eval
* Noel - Add code to save dialogue interactions into a file with item’s data (when returned after the user requests them)
* Aishwarya - Discuss prompt experimentation
* Everybody - write the paper
* Dan - add images to the front-end
* Dan - send out instruction for running front-end

Paper sections

* Dataset, Intro, Literature Survey, Abstract (and stuff from previous report) - Ameya

## Prompt techniques experimentation - aishwarya

## Human Evaluation- Ana (dependant on running model on computer)

## Noel - Perform auto evaluation

## Describe architecture - Dan

## 11/22

Action items:

* Build the dataset with images
* Have draft of code written for the auto evaluation - Ameya
* Send out code for with and without intent classification by 11/22 - Dan

## 11/17

Notes:

* Have an explanation for why we’ere seeing our results
* Have evaluation solid
* Micro benchmarks are solid too
* Have visuals of how the user would interact with the system (UI mocks)
  + And flow chart

Most important things:

* Focus on a couple things and evaluate them thoroughly
* Big focus on experiment driven evaluation

## 11/9 - Team Meeting

To do:

* Modify code to incorporate intent classifier model - Noel
* Make follow-up questions work (Dan)
* Fine-tuning results - Aishwarya
* Evaluation plan - Ameya and Ana

## 11/4 - Meeting with Bhavuk About Langchain

Objectives:

* Get his take on how to handle the different use cases
  + With an intent classification model
  + Without (i.e. natively in langchain)
* Understand why the event suggestions use case isn’t working well

Notes:

* CSV agent
  + Just works for direct query that I mentioned
* He thinks we should do
* Don’t
* One way that he’s doing it is using a master agent to choose which subagent to use
  + You can build a custom tool that just invokes the subagent
  + The master agent uses a custom tool
  + Use a tool decorator on a function to use a tool
* Feed a prompt into the master agent describing the two subagents which would be tool (found better results doing this way)

Questions:

* Why is context wrong after first question in output?

## 11/3 Clement

To ask clement:

* Is there a better way to do intent classification
  + He thinks the intent classification is reasonable but **we should experiment with and without it and include that in the report**
* Should we be doing fine-tuning?
  + Experiment with and without this too

Notes:

* Fine-tuning dataset is in format of query and response
* He doesn’t think it’s critical to have a numerical metric to evaluate on but it would be nice
  + One way to do this: pass outputs of event based suggestions to gpt 4 and ask if the items are relevant to {event from original prompt}
* Possibly measure follow-up queries too, (or evaluate anecdotally)
* Examples from fine-tuning data
  + {"prompt":"Suggest some food items for a traditional Filipino Noche Buena (Christmas Eve) celebration.","response":"1. Lechon (Roast Pig)\n2. Pancit (Noodles)\n3. Adobo (Chicken or Pork in Soy Sauce and Vinegar)\n4. Bibingka (Rice Cake)\n5. Lumpia (Spring Rolls)\n6. Halo-Halo (Mixed Dessert)\n7. Puto Bumbong (Purple Rice Dessert)\n8. Kare-Kare (Oxtail Stew)\n9. Sinigang (Sour Soup)\n10. Leche Flan"}

Action Items:

## 11/2

Notes:

* Won’t work on dataset until we’re getting it working on the dataset
* Maybe we should look into pretrained intent classifiers

## 10/30

* Preliminary results
  + Intent classification
  + Data set picture
  + Picture of the use cases
  + Fine-tuning
* Discussion
* paper
  + Ameya - abstract and data and vector search

## 10/26

Action items:

* Making the “event suggestions” use case work - Aishwarya
* Modifying the streamlit front-end to be acceptable for tuesday’s demo - Ana
* Becoming an expert in Faiss vectorDB - Ameya
  + Maybe LanceDB
* Intent classification - Noel
  + Understanding user intent

Notes:

* Performs best with a simple dataset
* What we can is breaking down prompts into multiple
  + E.g. “i need suggestions for thanksgiving dinner”

## 10/13

Notes:

* His thoughts:
  + It might be hard to force a database into vector embeddings
  + Thinks the vector database will be harder to set up but more precise down the long run,
* Eval
  + He doens’t think it’s clear how we’ll do evaluation
  + Suggestion: Evaluate each LLM system separately, then wholistically
  + Adding some objective eval metric:
  + For purely qualitative evaluation
  + He thinks we should compare to the instacart
* User interface: as long as it’s functional and you’ve done some good quality technical stuff then he’s not worried about it too much. even if it’s rows of a table
* Logistics
  + He doesn’t care how many times we meet. Willing to meet regularly but ok with 0 meetings

## 10/12

Objectives:

* Aggregate our findings
* FAISS
* Aishwarya plan for first milestone:
  + Says langchain can be used for the whole thing
  + Langchain with additional python libraries can be used to connect these 3 things
    - Database
    - Embeddings
    - LLM

Notes:

* MongoDB atlas vector search - hybrid search

Open questions:

* What is the lucene query used fun

## 10/5

Objectives:

* Decide whether to set up a meeting with the TA
* Divide work for next steps, get clear action items for each person
  + Finding a plan for getting the LLM to interact with the system. How do similar projects do it? Can we use Webarena? Is the data all there?

Action items:

* Set up a meeting with the TA
* Everyone come up with 5 examples of user interactions by tomorrow at noon
* Each person comes up with a recommendation for a specific path forward their below item by next Thurdsay:
* **Noel:** Recommendation process. how are existing similar systems doing this? What’s the best path forward
* **Dan:** Finding a plan for getting the LLM to interact with the system/how does it use data? How do similar projects do it? Can we use Webarena? Is the data all there?
  + Need to be able to have certain metadata about items (e.g. vegan friendly)
* **Ameya:** How to handle strict constraints (e.g. numerical, dietary restrictions)
* **Ana:** How will the natural language query be transformed into a structured/useful format (e.g. database query)?
* **Aishwarya:** Figure out how we can use langchain/other similar frameworks for this project. Where can it be valuable/ where can’t it?