Each week, add a new block above the existing blocks and fill it out by Sunday (that is, at the end of the quarter, the document should have Week 10, then Week 9, then Week 8, …). Each Sunday, submit a PDF of your doc as-is to Gradescope. Replace XXX in the title with your group number (e.g. A16-1). See here for more instructions.

Week 2 (due Sunday, January 19th)

Outside of your meeting with your mentor, when did your group meet this week?
Time: 5 PM, Saturday, January 18th
Location: Zoom
Attendees: Phu Dang, Eric Chen, Daniel Li, Bofu Zou

What did each group member do this week?

- Eric: During this week, I looked into the appliances in our smart home dataset, specifically the fridge, microwave, and dishwasher. I crafted 5 prompts and expected answers based on the ground truth data, and found out that microwaves and dishwashers have very similar monthly energy consumption patterns, whereas the fridge has a more seasonal curve to it (energy consumption is highest during the summer).
- Daniel: So this week, I focused on EDA in terms of the weather with respect to energy usage, as well as came up with some prompts related to those parameters. The main thing that I found was that most of the weather metrics had barely any correlation to energy consumption, if at all. The only two metrics that had any real correlation was temperature and humidity and those barely had much correlation.
- Bofu: During this week, I did a data analysis on the location part of the dataset, focusing on the trend and ground truth statistics of each location's contribution to the energy consumption. From the analysis, I found that the home office has consumed the highest amount of energy in 11 out of 12 months in a year, following with the barn. The change doesn't really follow the pattern of time. I tried to find the correlation between each location's energy use and temperature, but the correlation is relatively little.
- Phu: I worked on further experimenting with ways to effectively get the LLM to understand our tabular IoT data without using excessive tokens per prompt, as cost may be incurred fast. Since our smart home dataset contains timestamps in 1-minute intervals, the granularity of our data means we cannot include the entire dataset in our prompt, so I experimented with different aggregation methods. Besides a monthly group-by, I tried incorporating trends and summary statistics for each location and appliance in the home, which showed little improvements from our initial tests. I also set up our NeMo guardrails framework, the configuration files, and walked my teammates over the NeMo framework architecture for those who did not use NeMo in their individual projects last quarter. To incorporate energy prices data later, I picked a location in the U.S. that could have the sort of temperature range from our dataset, which is Denver, Colorado.

Where did they get stuck?

- Eric: Our development strategy for creating the prompts involved a step in which we were supposed to create a fact check guardrail. I wasn't too familiar with this since it was framed in the context of NeMo guardrails which I'm not very familiar with. Therefore, a lot of those sections were just me talking about how I would tell the guardrail to fact check the expected answer.

- Daniel: As I mentioned above, weather didn't have any real correlation to energy consumption, so I had to really dig deep to figure out if there was any real relationship between the two. Luckily, there were seasonal energy consumption changes that were shown in the weather fluctuations, so I went off of that.

- Bofu: I tried to dig out useful information and statistics about location to help build the chatbot, but it is hard to find out some patterns. I also have some confusion about the development strategy, more specifically, to come up with testing scenarios, but eventually I worked that out after talking with Eric and Phu. I am not really familiar with fine-tuning LLM, so it would be a challenge next week.

- Phu: I got stuck for a while at envisioning the holistic picture of our project and what our final deliverable will look like beyond the course requirements (paper, website, poster). We named our tool "Digital Bouncers" because we were excited at the idea of creating secure gatekeepers of digital interactions, a metaphor that allows our audience to relate an image of strength and controllability over their data. Like a bouncer at a club, our tool ensures that only safe, appropriate, and necessary data is allowed in or out. It filters and moderates inputs (user queries) and outputs (chatbot responses), blocking anything that doesn't comply with privacy, security, or appropriateness guidelines. However, perhaps because we were pushed to narrow the scope of our ideas, which we appreciate for easy planning and execution in the past 1-2 weeks, we inadvertently diverted from the security component and "bouncers" theme, which is critical for creating use cases for guardrails. For example, our two specific use cases: energy use summary creation (Spotify-wrapped style) and utility bill explanation, are too narrow for implementing guardrails as there are seldom opportunities for leakage of private information, which is the key demand driver for guardrails today, considering the common queries a typical homeowner may ask and the responses they receive (e.g., not many people would include their address when asking about a charge on their bill). Of course, a scenario we thought about is one of the homeowner's children may mess around, in typical children fashion, and ask our tool to say something inappropriate out of curiosity to see the tool's response; which would be a use case for guardrails, however, such scenarios are not compelling as they are rare. Nonetheless, exciting opportunities remain with RAG evaluation and output moderation guardrails (fact- and appropriateness-checking). We are confident that our

project is further "crystallizing" every week and we will figure out compelling use cases to have a novel and exciting project to serve our initial inspiration of *protecting and moderating digital interactions* with digital bouncers.

What will each group member do next week?

- Eric: Along with many of the other members of my team, we are moving forward with training and LLM on the smart home dataset instead of uploading raw data each time. More specifically, I'm going to figure out how to tokenize the tabular dataset so that it is ingestible for an LLM.
- Daniel: Next week, I will be experimenting with fine tuning the RAG with our specific dataset. I'll also get started on thinking about the streamlit app, and how we want to summarize the energy consumption breakdown for our app.
- Bofu: Based on the discussion this week, I think each of us would do some experiments with fine-tuning the LLM for our chatbot based on the dataset we have studied. It is also set to be a task that we need to tokenize the dataset. It is still on the agenda that I need to start working on the RAG development for the upcoming weeks.
- Phu: We will experiment with fine-tuning a pre-trained model, which I reckon the tokenization process will be challenging for our numeric IoT time-series dataset as LLMs are best for text data, given that they operate by predicting the next token/sequence of text. However, we believe it's a worthwhile experiment due to its unpopular nature and time-series data can be reformatted into sentences. For example: the following csv string "2025-1-19:17:15,0.09,0.01,0.05" with columns "timestamp,fridge,microwave,dishwasher" can be written as "The fridge used 0.09 kW, the microwave used 0.01 kW, and the dishwasher used 0.05 kW of energy at 17:15 on January 19, 2025". As I was writing this, I thought of splitting our management into two approaches to give the best shot at building a framework that can process time-series data reasonably well, so we might have two folks focusing on "fine-tuning" a pre-trained model, which we are unsure if the customized model would be compatible with NeMo, and the other two folks will focus on "in-prompt data provision" (as we have been experimenting) with LangGraph's built-in persistence layer to facilitate conversations.

  This will help "diversify" our approaches while keeping our "narrow" purpose of "processing time-series data."

  I will bring this thought back to my team.

As a group, do you still feel on track to execute your project proposal?
Yes. Despite the challenges above, we're confident that we're heading in the right direction and will create a novel and compelling case study of how LLMs handle time-series data with guardrails in the flow.

Schedule progression assessment — please see our updated schedule below:

Adjusted schedule (Spring 2025)

- Week 1 (DONE)
- Week 2 (DONE)
  - test prompts created
  - RAG documents identified (include in prompts directly (Zayd's demo))
  - final Streamlit chatbot deliverable pictured
  - attempted in-prompt data feed → has a sense of what works and doesn't
  - Settled on NeMo as the primary framework (covered configuration files architecture as a team)
- Week 3 (Jan 18 - 24)

```
- Experiment with fine-tuning (https://github.com/pndang/llm-comet/blob/main/fine_tuning.ipynb)
- Identify pre-trained model starting checkpoint (e.g., flan-t5-base)
- Use AutoTokenizer + Seq2SeqLM trainer for fine-tuning
- Use either Comet (already has working demo notebook) or TruLens for model registration/storage
- Steps: tokenize our tabular dataset, store in vectors, split into train and eval chunks, and train
```

- Week 4 (Jan 27 - Feb 1)

    - Begin integration into final framework (NeMo)
    - Begin RAG
    - Start testing input/output moderation and RAG eval test prompts

- Week 5 (Feb 3 - 7)
    - List key criteria for work distribution to begin drafting report materials
    - Split into poster- and web-team
    - Start drafting reports

- Week 6 (Feb 10 - 16)
    - Work on LaTex paper as a team

- Week 7 and onwards (contingency for progress overruns)

Week 1 (due Sunday, January 12th)

Outside of your meeting with your mentor, when did your group meet this week?
Time: 5 PM, Saturday, January 11th
Location: Zoom
Attendees: Phu Dang, Eric Chen, Daniel Li, Bofu Zou

What did each group member do this week? Where did they get stuck?

- Daniel: This week was mostly the logistic side of things. I went over the dataset very briefly to get a feel for it, and also had a better understanding of my specific role for the project. We talked about the general structure of our project and also split up the specific tasks that we'd each be working on.

- Eric: Since we mostly reviewed things we needed to work on this week, I helped with narrowing down to the scope alongside my other teammates before our first meeting this quarter. In our other meeting, we solidified a good plan of our project as well as who would be responsible for different parts of the data we are using for the project.

- Phu: I spent time exploring the dataset to get a good grasp of what's available for us to develop specific use cases and test scenarios (guardrails) for. My EDA work comprises of

graphs, summary statistics, figuring out what some ambiguous columns might mean, for example, we just found out that column "furnace" is the sum energy consumption for both furnaces in the subject house (labeled "furnace_1" and "furnace_2" — I think little insight nuggets like this are crucial for us to effectively and accurately work with the data we have. I also tested out how well we could get the LLM to digest and process tabular data, which we are keeping track of good prompts + responses to potentially use as one-shot and/or few-shot context learning later. Our current observations are heading in the right direction.

- Bofu: For this week, in general we discussed the content and workflow of our project and set up a plan throughout the entire quarter. I downloaded the dataset and got to know the basic structure and figures of each column and picked my job in the team plan.

What will each group member do next week?

- Daniel: I will be doing some data analysis on the effects of weather on the electricity bill. I will also be working with Bofu on the functionality of delivering the summary on electric usage. In terms of the guardrails, I will explore different options within Nemo guardrails to see which ones we could apply to our LLM. Finally, since my ESA is on weather, I will be coming up with weather related test prompts to see the performance of our LLM on that.

- Eric: My main task is to understand appliance data in the smart home dataset, particularly trends and ground truth statistics, and ideate specific prompts and expected (correct) responses. But I also will be working with Phu on finding utility prices for our dataset because alongside energy savings, we want to think about how we can lower costs for people since this is about smart homes and how we can build tools for a better future.

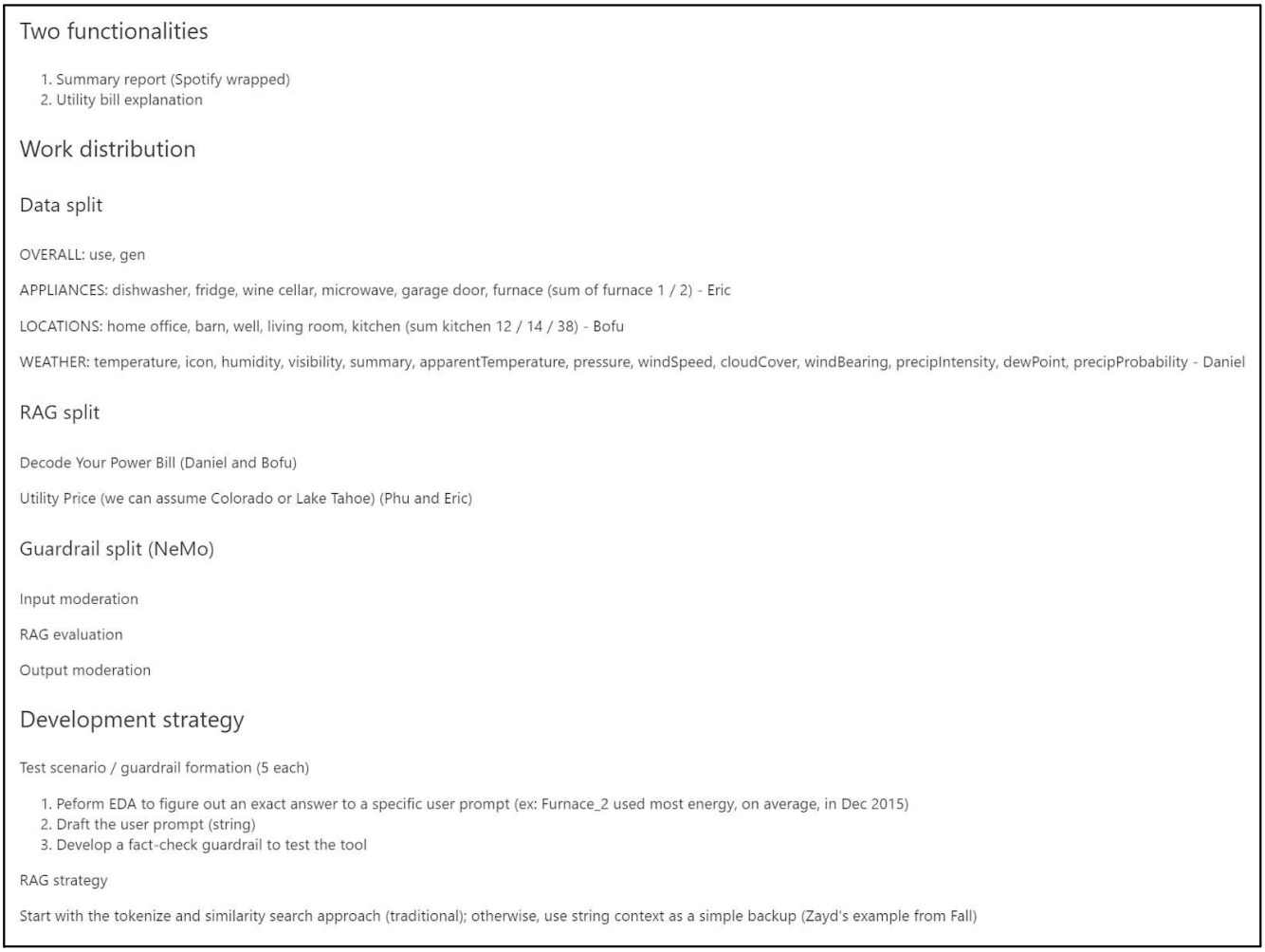
- Phu: I will conduct prompt engineering to best instruct the LLM in understanding its role as a smart home data assistant, this would be the "system" content prompt portion of an OpenAI API call, or the role prompting in NeMo guardrails' Colang configuration files. The other end would be response formatting; since this project was inspired to be an AI smart home product proof-of-concept, I will include instructions for the LLM to return user-friendly responses to keep length, wording, conversational flow in check for the best user experience.

- Bofu: My task for next week is about doing analysis on the location part of the dataset. I will analyze how each location contributes to the power bill and how each location's electricity usage will vary with other variables (like weather). I will also find out the difference between each location and how the difference makes changes on the bill. For the RAG split, I will cooperate with Daniel to work through decoding the bill.

As a group, do you still feel on track to execute your project proposal?

Yes, we are confident that we are on track to execute this project, largely thanks to the "specificity" that our mentors pushed for us to achieve in our proposal, which helped us work very conveniently by thinking of concrete tasks and goals to achieve each week, at least in Week 1 for now, we have a clear vision of the final product, what to be done, and work distribution. We think one of the most productive plans that could turn out to be very beneficial to us by the end of this quarter is "scoping" (screenshot below). For example, we have decided on two specific RAG documents, serving two specific functionalities, that is straightforward to develop guardrails / test prompts for. And to facilitate this testing, we developed a concise 3-step process to determine the ground truth to be tested against through EDA, then one specific prompt / question a user may ask, and measure accuracy as components of a fact-check guardrail framework.

Screenshot from explore.ipynb



Two functionalities

1. Summary report (Spotify wrapped)
2. Utility bill explanation

Work distribution

Data split

OVERALL: use, gen

APPLIANCES: dishwasher, fridge, wine cellar, microwave, garage door, furnace (sum of furnace 1 / 2) - Eric

LOCATIONS: home office, barn, well, living room, kitchen (sum kitchen 12 / 14 / 38) - Bofu

WEATHER: temperature, icon, humidity, visibility, summary, apparentTemperature, pressure, windSpeed, cloudCover, windBearing, precipIntensity, dewPoint, precipProbability - Daniel

RAG split

Decode Your Power Bill (Daniel and Bofu)

Utility Price (we can assume Colorado or Lake Tahoe) (Phu and Eric)

Guardrail split (NeMo)

Input moderation

RAG evaluation

Output moderation

Development strategy

Test scenario / guardrail formation (5 each)

1. Peform EDA to figure out an exact answer to a specific user prompt (ex: Furnace_2 used most energy, on average, in Dec 2015)
2. Draft the user prompt (string)
3. Develop a fact-check guardrail to test the tool

RAG strategy

Start with the tokenize and similarity search approach (traditional); otherwise, use string context as a simple backup (Zayd's example from Fall)