Prompt Engineering & Evaluation Assessment

TIM 175 WEEK 6 ASSESSMENT

This is a small mini-project assessment task. There are **TWO** components to this: 1) your initial work and thoughts that you will **submit at the end of the class,** and 2) a take-home component in which you will continue improving and expanding on your initial work as an assignment **due by Thursday before class**. If you need an extension on the take-home component, we can be flexible if you want to continue working on it until Saturday 11:59pm (no further extension beyond that).

For the initial submission at the end of class, treat it like a talk-aloud interview where you describe your thought process for how you are thinking about the problem as you begin to make progress on it.

You need to create a full prompt engineering workflow to tackle the problem given. Here are some guidelines:

- Take some time to **understand the problem**. Make sure to document your thoughts so we can understand your thought process. You don't need to have everything figured out before you get started
- As you work on the problem, you should make sure to touch on each of the following aspects:
    - **Creating a workflow: s**how us how you think about when to break down the problem into smaller tasks whose output feeds into the next,
    - **Designing and iterating on prompts**: create prompts and iterate on them using prompt engineering techniques with an eye towards producing high-quality outcomes,
    - **Evaluation:** show us how you evaluate outputs and define metrics that can evaluate the outputs. In the take home when there is more time, you should also show us how you can operationalize metrics in prompts using LLM-as-a-judge,

Submission Instructions:
Week 6 Assessment Submission

You will need to submit:

- A Google Document that briefly documents your thoughts. You will submit a link to this document and also copy-paste the document's contents into the Google Submission Form. Try to make sections in it for easy readability (e.g initial thoughts, Initial Prompt, Refining, Evaluation).
    - For the end-of-class submission, we expect you to:
        - briefly outlines your initial thoughts about how you might break down the task,
        - defines and runs at least one prompt, and discusses how you plan to iterate on it next based on results,
        - share initial thoughts on evaluation metrics you will use.
    - For the final submission, we expect you to:
        - share the process of how you ended up breaking down the test and what you learned from making your own prompt workflow

    - Document how you iterated on the prompts over time to make them high-quality based on the initial outputs
    - Explain how you conducted manual and LLM evaluations and what you learned from it.
- A LastMileAI Workbook or Google Collab Notebook of your choosing for creating your Prompt Engineering Workflow. Both are equally good to us for the purpose of this assignment. Here is the Colab template you used in Week 4.

You can use ChatGPT or other GenAI tools to inform any part of the assignment but: (1) you need to first form your own independent thoughts, (2) every word included in the submission needs to be something you've read, thought about, and decided to include, and (3) you should strive towards submitting the highest quality work you can rather than mediocre work that meets the requirements.

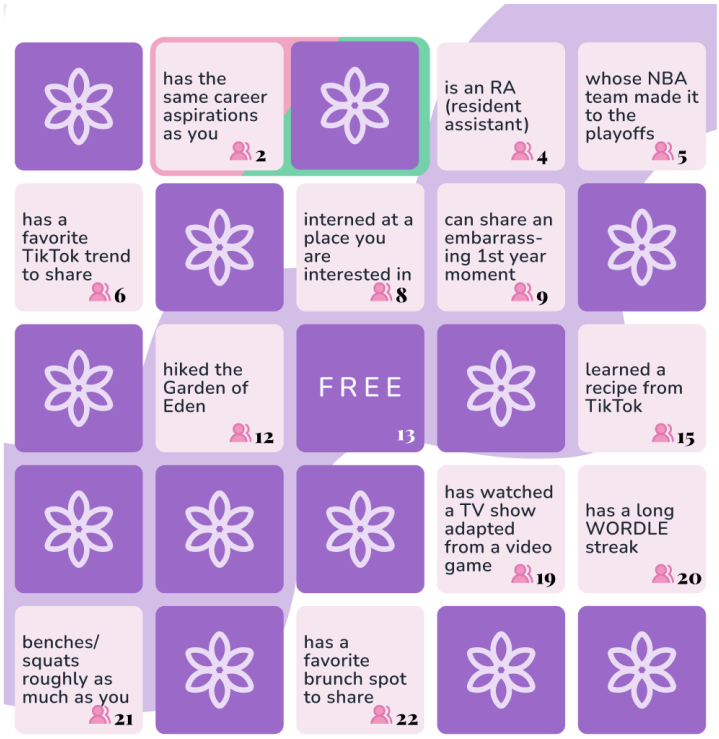
Personalized Human Bingo

In today's task, you will use LLMs to create personalized Bingo Squares to facilitate connection in a Human Bingo Activity (a simplified version of the personalized Human Bingo activity you participated in during the first section at the beginning of this quarter).

Human Bingo: Human Bingo is an icebreaker activity that helps to facilitate connection in an in-person community event. Traditionally, each individual is given a 5x5 Bingo Board with Bingo Squares containing fun random descriptors about attendees (e.g. "never seen star wars", "loves boba", etc.). Attendees mingle around, introduce themselves to new people and search for those who can help them check off a bingo square containing a matching descriptor (max one square per person).

Personalizing Human Bingo for Connection: We'd like to personalize the bingo squares so that they aren't just fun random facts, but also contain squares that facilitate more meaningful connections between attendees over shared interests, goals and hobbies. For example, if one participant loves sports, it might be particularly helpful to connect them with someone else who loves sports. If two people want to do startups in edtech, it'd be great if we could help them find each other.

Below, you can see an example of a human bingo board we made in the past. Some of the squares on this board are intended to help facilitate connection and are designed based on participant responses to a survey about their interests, goals, hobbies and more.



Participant Data: Before the event takes place, people attending the event fill out a pre-event survey about their interests and goals. We'd like to design LLM prompts that help to generate bingo squares (and ultimately, bingo boards) for each participant based on the survey responses. For your task, you can use the following dataset containing real survey responses from a past event we ran: Participant Data - Live Task.

Task: Create a Prompt Engineering Workflow including a series of prompts and evaluations that use the participant data given above to generate Bingo Squares that are likely to facilitate connection for the participants. We will be generating a subset of the squares (roughly 7-9) on the bingo board in this task:

1) 6 squares in each bingo board are those that are **general to everyone** and should represent themes extracted from **the responses of all participants.**

    a) Of these, 3 of these should be topics that <u>many participants</u> have expressed interest in and will help make connections over the majority shared interests.

    b) While the other 3 should be topics that <u>only 1-2 other participants</u> expressed interest in to make those squares niche and more difficult to achieve.

    c) You also want there to be diversity in the topics represented in these six generated squares.

d) Your prompt for this part should output, for each of the 6 squares generated, the bingo square text and a list of the participants (i.e. Participants N, Q, S) whose responses show a match with that bingo square,

e) ***Tip: To tackle this problem, we recommend you have the LLM first extract major themes in the participant responses and group the responses into those themes (one response can be in multiple themes). Then have it create the Bingo Squares from the majority and minority themes.***

2) Up to 3 squares in each bingo board should be **personalized to each participant.** In other words, your prompt should take ONE given participant as input (along with the set of all responses) and it should produce up to 3 bingo squares that relate to the given participants' interests to add specifically to their board.
    a) Your prompt for this part should output, for each of the 3 squares generated, the bingo square text and a list of the other participants (i.e. Participants N, Q, S) whose responses also match with that bingo square (have a shared interest),

    b) If the participant did not share 3 interests in their survey, then you should only output a number of squares corresponding to the number of interests they shared. ***Note:*** *even if an interest they listed does not match anyone else's interests, it is still a candidate for being made into a bingo square. So if someone shared 2 interests in their survey, there should be 2 bingo squares generated even if neither of those 2 interests are shared with others.*

    c) If the participant shared more than 3 interests, you should pick the ones that are most represented among other participants. For example, if one interest is also shared with 5 other people, you should prioritize including that over an interest shared with only 1 other person.

    d) ***To tackle this problem, choose 1-3 participants first and then create/refine the prompt for making personalized squares. Repeat with other participants to test it.***