TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

Evaluation Metrics, Iteration, and Dataset Curation

TIM 175 WEEK 3 PRELAB

Today, we will learn to define metrics for rigorously evaluating AI outputs and to use these metrics for further iteration on prompts. We'll focus on how to measure and improve prompt effectiveness through systematic evaluation and refinement. **This individual prelab is due Tuesday 11:59pm.**

Readings: We mark up to two readings with a ⭐that we suggest you read

- AI Alliance NYC Meetup -- LastMile AutoEval ⭐
- LLM-as-a-judge: a complete guide to using LLMs for evaluations ⭐
- Deploying an application with Generative AI best practices
- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org
- GenAIOps: Operationalize Generative AI

Submission Link
Week 3 TIM 175 Submission Form (Spring 2025)

Brief Task Overview
Set up LastMile AI and create a copy of this document to work through the activities. For each activity:

- Capture screenshots of the prompts you used and the outputs generated by the model. Place these images directly **above** the analysis tables of each activity.
    - No screenshot is needed for Activity 1 (manual evaluation) and Activity 4 (which is focused on dataset curation)
    - For Activity 2-3, two screenshots are sufficient (you do not need to screenshot every single user query + criteria)
- Use the table to write a detailed analysis of your reflections from the activity.
- ALL NEW TEXT YOU ADD SHOULD BE IN RED TO MAKE IT EASY FOR US TO SEE.
- Create a Google sheets spreadsheet for documenting your prompt refinements using the following template: 3. Annotated Spreadsheet

Complete the following activities to understand methods for evaluating prompts and outputs:

1. Manually Evaluating Responses with Metrics
2. LLMs as a judge/ LLM as evaluators
3. Pairwise Comparison with LLM-as-a-judge
4. Creating an Annotated Dataset During Iteration
5. (optional) Experiment with Temperature

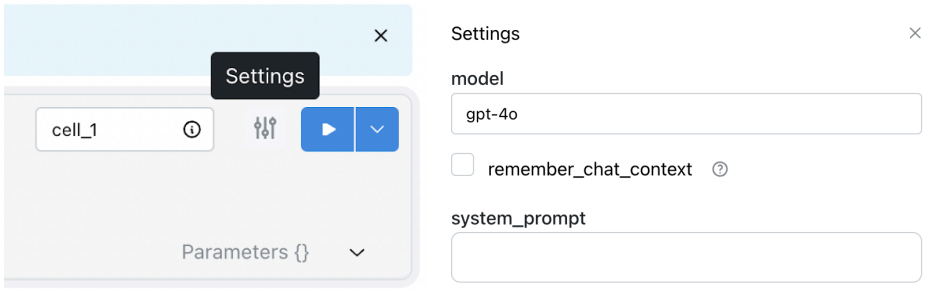
You can use ChatGPT or other GenAI tools to inform any part of the assignment but: (1) you need to first form your own independent thoughts, (2) every word included in the submission needs to be something you've read, thought about, and decided to include, and (3) you should strive towards submitting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

TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

Setting up

Set Up LastMileAI:
To try out different prompt engineering techniques we will be using the following platform: LastMileAI.

1. Simply Sign in to the platform using your Google Account
2. Click on Workbooks and then "New Workbook"
3. You can experiment with different models using the dropdown on the top left. By default it is set to "ChatGPT". For this assignment, **please change it to GPT-4**.
4. **IMPORTANT:** For this activity, go to settings on the workbook cell and deselect "remember_chat_context"



Evaluation Metrics and Iteration Processes

Recall that prompt engineering is the process of designing and optimizing prompts to guide AI models to generate the desired outcomes. Remember, a model's output can only be as good as your input is - if you offer the AI a poor prompt, you can limit the quality of its response. To effectively iterate on prompts, we need to have proper metrics for evaluating our outputs so that we can systematically improve our prompts.

Key Evaluation Metrics: Some commonly used metrics include:

- **Relevance:** Does the output align with the prompt?
- **Coherence:** Is the response logically structured and clear?
- **Completeness:** Does it address all aspects of the prompt?
- **Fluency**: Is the language natural and well-formed?
- **Diversity:** Does it provide varied and creative responses when appropriate?

These are not the only metrics that are used. For example, when context is provided, Faithfulness and Context Utilization are two other common metrics. Additionally, you will almost always need to define your own evaluation metrics tailored to your specific task context.

TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

Activity 1: Manually Evaluating Responses with Metrics

Objective: Explore how to manually evaluate different AI outputs for some of the key evaluation metrics.

Task 1: An Airline company has incorporated a customer-support LLM chatbot into its website. The table includes queries asked by the customers and the answers from the chatbot. Analyze each of the outputs and fill the metrics box by giving it a yes or no label. You will notice how we add new metrics based on the context.

- **Relevance**: Does the response align with the user query?
- **Completeness**: Does it address all aspects of the user query?
- **Helpful:** Is the response helpful to the user?
- **Coherence**: Is the response logically structured and clear?
- **Tone:** Is the response polite and encouraging as wanted for customer support?

Complete this table in **your copy of the Google Spreadsheet** in the tab called Airline Activity and add manual evaluations scores and reasonings for the scores (Column D to I).

User Query	**Output**	**Scores (yes/no):**	**Reasoning for your scoring? How could the output have been better?**
Can I bring my pet on the flight? What rules do I need to follow?	Yes, pets are most welcome on our flights. You'll need to pay a small fee and follow our pet policy guidelines which can be found on our website.	**Relevant?** **Complete?** **Helpful?** **Coherent?** **Tone?**	
I can't find my passport, what should I do?	I am sorry about that. Passengers are required to present a real form of identification to fly.		
Where do I check in for international flights?	For international flights, you can check in at the main terminal, counters 5 through 10.		
Can I change the name on my ticket after booking?	No you cannot.		
What beverages are	Yes, You can make a		

available on the flight? Can I get a vegetarian meal on the flight.	request for a vegetarian meal on the app when you check-in. We love to accommodate all requests of our passengers to the best of our ability.		
How long is the flight from Chicago to Los Angeles? What is the time zone difference between the two cities	6 hours.		
Can I charge my laptop on the flight?	Our airplanes have free Wifi access on board along with an entertainment system with a variety of movies and tv shows.		

Reflection: What did you learn about different metrics? What other metrics do you think the customer support bot should be evaluated on?

Your reflections:

TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

Activity 2: LLMs as a judge/ LLM as evaluators:

Objective: In this activity, we are going to explore using LLM-as-a-judge for evaluating responses and compare them to manual evaluations.

Required Reading: LLM-as-a-judge: a complete guide to using LLMs for evaluations

Make sure that you thoroughly understand the reading - **It is very important that you read it before continuing so that you understand.**

LLM-as-a-Judge is an approach in which LLMs are used to evaluate AI-generated outputs based on an **evaluation prompt** that defines **evaluation criteria** for what different metrics mean, i.e. what it means for an output to be relevant or to have the desired tone**.** Once you have done so, you can take the **text output** from your AI system and feed it back into the LLM to obtain a score, label, or even a descriptive judgment—following your instructions.

Before going forward, make sure you understand the following:

- What is an evaluation prompt? What are the important elements in it?
- How do you define a custom criteria / what should the criteria include?
- Why is LLM evaluation useful? When does it work well and when does it not?

Task 1: For the same customer support activity from Activity 1, we will now evaluate the responses using an LLM. You can use the same LastMileAI workbook to run your LLM evaluations, selecting the same foundation model (GPT-4).

Open LastMileAI and do the following:

1) Create an evaluation prompt defining evaluation criteria that explain each of the 5 given metrics (Relevance, Completeness, Helpfulness, Coherence, and Tone).
2) Apply your evaluation prompts to each of the input-output pairs to obtain an evaluation and explanation

Your Evaluation Prompt

Complete the table in **your copy of the Google Spreadsheet** in the same tab called Airline Activity by filling out **Column K to P** with the evaluation provided by using LLM-as-a-judge and the reasoning behind each evaluation.

User Query	**Output**	**LLM evaluation scores (yes/no):**	**Reasoning for score given by LLM evaluation**
Can I bring my pet on the flight? What rules do I need to follow?	Yes, pets are most welcome on our flights. You'll need to pay a small fee and follow our pet policy guidelines which can be found on our website.	**Relevant?** **Complete?** **Helpful?** **Coherent?** **Tone?**	

Reflection: How do the scores and reasonings from the LLM differ from your manual evaluations? Do you think the LLM evaluated the queries properly for each of the metrics? Are there ways in which your evaluation criteria and prompt can be improved?

Your reflection on the comparison between manual and automated evaluations:

Based on your reflection, refine the evaluation prompt and criteria and then complete another iteration in the spreadsheet (Columns R to W). Reflect on the difference between evaluation iterations.

Refined Evaluation Prompt

Reflection on Refining Evaluation Prompt

TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

Activity 3: Pairwise Comparison with LLM-as-a-judge:

Pairwise comparison: You can also use LLM evaluators to compare two different LLM responses and ask it to choose the better one. This lets you compare models, prompts, or configurations to see which performs best.

Resource: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org

Task: You will need to use LLMs three different times for this task.

- You can use two different models in LastMile AI, e.g. gpt4 as LLM 1 and ChatGPT as LLM 2.
- Give LLM1 and LLM2 the same prompt telling it to compose an engaging travel social media post about a recent trip to Italy, highlighting cultural experiences and attractions that viewers must visit. The post should not be longer than 5-6 lines and should be in an engaging and friendly tone. Use the exact same prompt for both of LLMs.
- Use an LLM a third time, this time for evaluation to compare both outputs and choose which one is better.
- You will need to create an evaluation prompt that will take the prompt, the outputs from LLM1 and LLM2, and instructions on how to evaluate the outputs (a criteria). The LLM should choose one and give detailed reasons for the choice.

Complete the Travel Social Media Row in **the same copy of the Google Spreadsheet** in the tab called **Pairwise Comparison** by filling out the two original LLM responses and the LLM evaluation output.

Task 2: Repeat the same activity as above with a problem context that you think of yourself and create a prompt for yourself. Add a new row in the same "Pairwise Comparison" tab for your problem context.

Reflection: Why did you choose this specific problem context? How were the outputs between LLM1 and LLM2 different? Do you agree with the LLM evaluation's choice and reasoning? Are there any ways in which you would want to refine your evaluation prompt?

Your reflections:

Activity 4: Creating an Annotated Dataset During Iteration:

Evaluation and iteration. Evaluation is critical to informing further iterations of your prompt design. You may notice that your outputs are good for one metric but are lacking for some specific metric. In such cases, you need to think of ways how your prompt could be refined to improve the output for specific metrics. For example,

- If **answer relevancy** is low, adjust your prompt to be more explicit about the focus of the answer and problem context.
- If **completeness** is low, make sure your prompt lists out all the required aspects of the task.
- If **format or tone** is not correct, make sure your prompt specifies exactly what format/tone is required through instructions, examples or role assignment.
- If **faithfulness scores** are low, consider if your prompt needs more guidance to ensure the response stays aligned with the source.
- If **context utilization** is low, think about whether your prompt could better specify which context elements to draw on.

You may also consider these other tips on how to refine your initial prompts:

- **Adding Specificity:** Review your initial prompts and identify areas for more detail. For example, instead of "Tell me about space," try "Describe the process of star formation in a nebula." You can also add more context or examples to improve specificity in your prompt.
- **Changing Structure:** Experiment with different formats, such as questions or statements. For instance, ask, "What are the key stages of star formation?" versus "List the key stages of star formation and explain each briefly."
- **Role Assignment:** In certain cases, telling the LLM to take on a specific role for a specific context can be valuable e.g "You are a science teacher that needs to explain the process of star formation to 5th graders. Think of an interactive and easy way to explain the key stages of star formation that will be engaging for students of that age."
- **Experimenting with different models:** In a typical scenario, you could also explore different AI models to assess their impact on output quality. However, for this class, we will use the GPT-4 option in LastMile AI.

Creating an annotated dataset. Evaluation is not only critical to do manually for informing your prompt design and iteration, it can also be used to systematically evaluate future prompts or to even create fine-tuned versions of the LLM or develop automated evaluator models. To do this, it's important to curate an annotated dataset throughout the process of manual iteration and evaluation.

TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

To develop an evaluator model for a given metric, we want to collect a set of inputs, possible generated outputs, and labels evaluating the generated output based on the metric criteria. Our evaluator models will be binary classifiers, so each data point should have the following

- **Inputs:** these are the unique parts in your prompt template, e.g. the user input or text you plug into your prompt,
- **Output**: a potential model generated output for those inputs, e.g. what was generated from one of your prior iterations,
- **Labels**: a 0 or 1 based on the metric evaluation criteria, e.g. whether the generated output was relevant or not relevant given the input and criteria for the relevancy metric,
    - **LLM evaluation labels**: we may store intermediate LLM-generated evaluations created using LLM-as-a-judge,
    - **Human evaluation labels**: these LLM evaluation labels are then verified/edited by humans manually to get the true/final labels.

It can also be helpful to curate other labels useful for fine-tuning the base model:

- **Ground truth**: the expected or ideal output you would like from the LLM, e.g. collected by experts writing responses that are a best attempt at the ideal response,
- **Human rank ordering**: a ranking of multiple generated outputs, used for Reinforcement Learning with Human Feedback (RLHF).

Task: Practice creating an annotated dataset while iterating on a prompt for the following problem:

- **Given a user response to an open-ended survey question, we would like to generate a follow-up question that elicits rich stories and experiences from the respondent like in an interview,**

The inputs to the prompt would consist of:

- **The survey question**, e.g. “How should UCSC improve the student experience?”
- **The survey response**, e.g. “Housing is horrible! Make that better”

You should define a prompt that generates a follow-up question that elicits richer stories and experiences.

- **A good example**: “Can you share a horrible experience you had related to housing?”
- **Bad example**: “What about housing is horrible?” (this does not ask about the respondents stories and experiences, which is the focus of this prompt)

As you iterate on your prompt, you should use the following custom metrics:

- **Elicits stories**: we want follow-up questions that are likely to elicit rich stories and experiences from the respondent,
- **Relevant and logical**: we want follow-up questions that are relevant and logical follow-up questions given the response,

TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

For this activity, use the third tab in your copy of the spreadsheet called "Annotated Dataset." Follow the following steps and complete the annotated dataset:

1. Define at least **three** different survey questions for different contexts related to getting feedback on a product, asking for suggestions/ideas for improvement, or understanding needs around a problem area. For each survey question, define **three** different survey responses. This gives you **nine** different inputs of survey questions and responses. Add these under columns A and B. Note: you may use the example input in the spreadsheet template.

2. Write a prompt that tries to generate an effective follow-up question that elicits rich stories and experiences from the respondent based on the input survey question and response.

3. Write an evaluation prompt for each of the two metrics (elicits stories, relevant and logical). The evaluation prompts should take the inputs and output and return a 1 or 0, with 1 representing a high-quality prompt meeting the metric criteria.

4. Test your prompt on each of your nine inputs **two** times to obtain two generated outputs per input. Record each generated output under a different row under column C.

5. Run your two evaluation prompts on each row (set of inputs and an output) to generate LLM evaluation labels for each metric. Record the generated evaluation labels under columns E and H.

6. Manually evaluate the LLM generated labels to determine the final manual labels under columns F and I, copying over LLM labels you deem to be correct and changing LLM labels you deem to be incorrect. Add a rationale for any LLM label that you deemed incorrect, and optionally, for LLM labels that you felt were correct.

7. Reflect on what is working or not working in your main prompt and your evaluation prompts. Make a new iteration to address any weaknesses and go through steps 4-6 again with your new set of prompts (except this time, in Step 4, you only need to generate one output rather than two). This will result in **three** rows of outputs for each input.

8. If none of the generated outputs are ideal, create a ground truth output (i.e. an ideal follow-up question) and add it as a separate row for the same inputs. This means there will be either 3 or 4 rows of outputs per input. Finally, rank these 3 or 4 possible outputs in column D, with a rank of 1 for the best output (e.g. the ground truth output), then 2 for the 2nd best, and so on.

Reflection: How did you improve your main prompt and evaluation prompts to improve the quality of the outputs and labels?

Your reflections:

Submission Instructions:

After completing all the above activities, fill out the submission form including a link to your copy of this google document on it and a link to your spreadsheet.

(OPTIONAL) Activity 5: Experiment with Temperature

The temperature parameter influences the randomness and creativity of generated outputs in AI models.

- Lower Temperatures (0.0 - 0.3):
  - **Characteristics:** Outputs are more deterministic and consistent.
  - **Use Cases:** Best for factual information or tasks that require clarity and reliability.
  - **Example:** A prompt like “What is the capital of France?” will consistently return “Paris.”
- Medium Temperatures (0.4 - 0.7):
  - **Characteristics:** A balance of creativity and coherence.
  - **Use Cases:** Good for creative writing and brainstorming where some variability is desirable.
  - **Example:** A prompt may yield relevant yet interesting ideas.
- Higher Temperatures (0.8 - 1.0+):
  - **Characteristics:** Outputs are diverse and creative but may lack coherence.
  - **Use Cases:** Ideal for artistic projects or generating novel concepts.
  - **Example:** A prompt for a “fantastical story” can lead to unique and unexpected narratives

The temperature parameter can either be manually adjusted in an environment like in the LastMileAI settings or by using certain words that let the LLM know within the prompt to yield different results. Example: For a prompt like: "Write a story about a robot and a cat." you can adjust the temperature by changing the words as below:

- **Low temperature:**
  - Prompt: “Provide a two-line, concise, realistic story between a robot and a cat”
  - The robot and the cat might have a straightforward adventure with predictable outcomes.
  - “A delivery robot paused as a stray cat climbed onto its warm top, seeking refuge from the cold night. Without hesitation, the robot adjusted its route, heading to the nearest animal shelter.”
- **High temperature:**

- Prompt: “Provide a two-line, fantastical story between a robot and a cat”
- The story might take unexpected turns, such as the robot and cat forming a rock band or traveling to another dimension.
- “A solar-powered robot built a tiny rocketship for a talking cat that claimed to be the ruler of the Moon made from a yarn. Together, they set off to reclaim the feline's celestial throne.”

Hence, by adjusting the temperature, you can tailor the model's output to suit your needs, balancing creativity and precision and resulting in different evaluation scores.

Example For a Blog Post Prompt:

- Low Temperature (0.0 - 0.3)
  - **Expected:** Clear, factual content with scientific references
- Medium Temperature (0.4 - 0.7)
  - **Expected**: Balanced personal insights and facts showing mix of creativity and facts]
- High Temperature (0.8 - 1.0)
  - **Expected**: More creative, varied perspectives showing unique angles

Task: Now define three prompts for the following problem context one for each: low, medium and high temperature:

Problem Context: You are part of the debate club's social media team, and your task is to create captions for multiple posts about an upcoming debate competition hosted by your school. The captions need to match the tone and style that different audiences prefer and you need to come up with three captions for each kind of preference to submit to the team.

- Low Creativity: Straightforward and informative, focusing on clear details.
- Medium Creativity: Engaging and slightly fun, with a touch of personality.
- High Creativity: Bold, imaginative, and attention-grabbing to spark excitement.

Tip: Use specific words in your prompts to guide the AI, such as “informative”, "simple," “straightforward”, "engaging," or "imaginative."

Create three versions of a prompt (one for each level of the temperature.) and complete the following table:

Temperature	**Prompt**	**Scores (1-5):**
Low		**Coherence:** **Creativity:** **Factual Accuracy:**
Medium		**Coherence:** **Creativity:**

		Factual Accuracy:
High		**Coherence:** **Creativity:** **Factual Accuracy:**

Reflection: What words or phrases in your prompts encouraged more creative scores? How did the tone of the captions change across low, medium, and high temperature? What did you learn from this?

Your reflections:

TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025