TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

Prompt Evaluation of Initial Coding

TIM 175 WEEK 3 LAB

In our first project (Weeks 2-5), we will use Generative AI to extract themes and create rich narratives around career insights grounded in the stories of professionals in the community shared on YFIOB's What-To-Be podcast. Last week you used prompt engineering techniques to identify relevant passages and create initial codes from the interview transcripts. This week's objective is to create rigorous evaluation metrics and prompts, and to curate a labeled dataset while you iterate further. This will include an **individual deliverable (due Saturday at 11:59pm)** and a **team deliverable (due Monday 11:59pm)**

Readings: see HW 3 Prelab readings on prompt evaluation.

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

Brief Task Overview
Individual deliverable

1. Create a workbook on LastMileAI
2. Update your main prompt from Lab 2
3. Define evaluation metrics, make a copy of this spreadsheet
4. Prepare the dataset curation tabs
5. Manually evaluate a few generated responses
6. Define an evaluator prompt for each evaluation metric
7. Run and refine your evaluator prompts.
8. Apply your final evaluator prompts to ALL data for the given transcript
9. Reflect and submit using this Google Form

Team deliverable

1. In-section peer review of evaluation
2. Compile your reviews, make another copy of this spreadsheet
3. Iterate as a team
4. Reflect and submit using this Google Form

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.

Individual Assignment Instructions

Create a workbook on LastMileAI. Go to https://lastmileai.dev, and create a new workbook titled “Project 1 - Week 3 Lab”. This is where you’ll be constructing your prompts. Remember to use the GPT-4 model.

Warning! Make sure you don’t exceed the weekly token limit in LastMileAI. This can happen if you change the models to a heavy model like TTS. So please stick to using GPT-4 or ChatGPT models and do not use LastMileAI for anything irrelevant from the Lab.

If you cannot use LastMileAI for any technical reason and you receive an error when you run a cell, try restarting LastMileAI. If it does not work, please inform a TA/tutor. You may then use this Google Colab as a template to generate the rest of your outputs.

Problem task: initial coding with evaluation and dataset curation. Like in Lab 2, you will be defining and iterating on prompts to do initial coding of transcript chunks for YFIOB’s What-To-Be podcast episodes. However, this time, you will also be doing rigorous evaluation as you refine your prompt and also curating an annotated dataset. For the individual portion of the assignment, you will be using transcript chunks from this single transcript:

- 078_Kayla Baumgardner Firefighter Paramedic

Update your main prompt. Update your prompt from Lab 2 individual or team that identifies relevant passages from a transcript question-answer chunk. You need to make sure that the prompt response ends with a set of quotes and initial codes formatted as an array of objects with include_code and quote properties as shown below:

```
[
  {
    initial_code: “<the code>”,
    quote: “<the quote>”,
  },
  {
    initial_code: “<the code>”,
    quote: “<the quote>”,
  },
  ...
]
```

Define evaluation metrics: Define a set of evaluation metrics that will capture all aspects of a high-quality output given an input transcript chunk. For example, your set of metrics should enable us to determine whether a given output:

- Correctly identified exactly the right set of relevant passages for the research question (did not miss any, did not include incorrect ones, etc.),
- For each item identified, has a clear well-phrased code/theme that captures what about the passage is relevant or might be relevant to the research question,

- For each item identified, has a quote that includes all relevant text for that code/theme, but does not include irrelevant text that would distract from the code/theme,

Some guidelines:

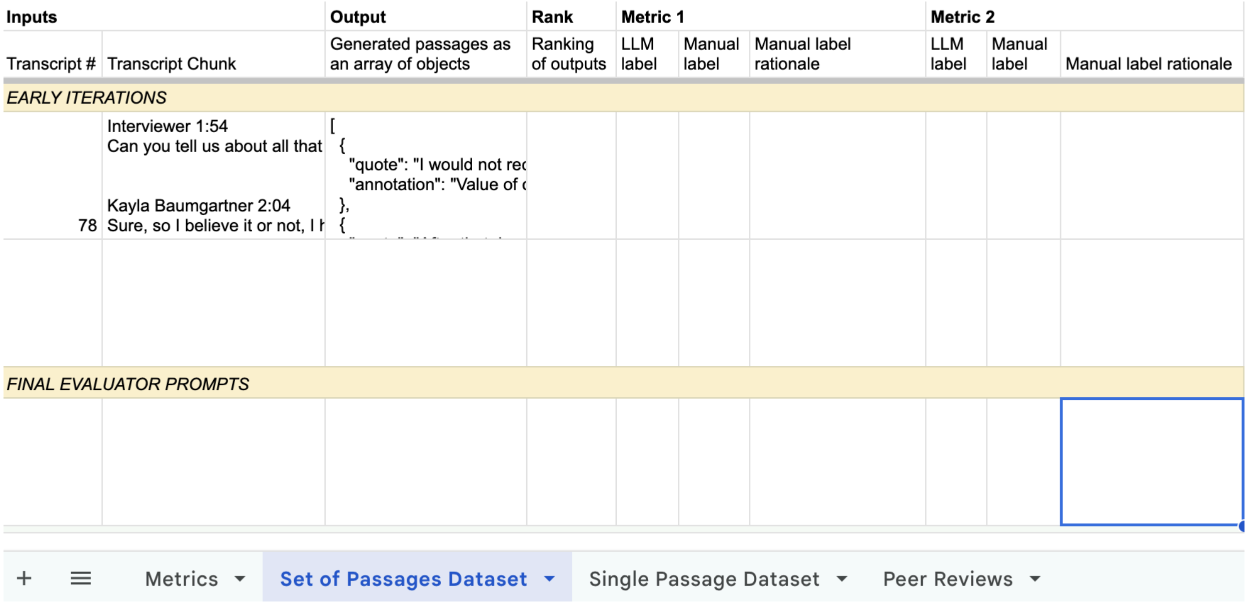
- For the first bullet point, you will want to define at least two separate evaluation metrics (e.g. whether you missed any, whether there are incorrect things included),
- For the second and third bullet points, your evaluation metrics should be applied to a single relevant passage in the output array rather than the entire set of passages,

Once you've defined a comprehensive set of metrics, make a copy of the annotation spreadsheet and add these metrics to the 'Metrics' tab. Each metric should have a name and a brief description.

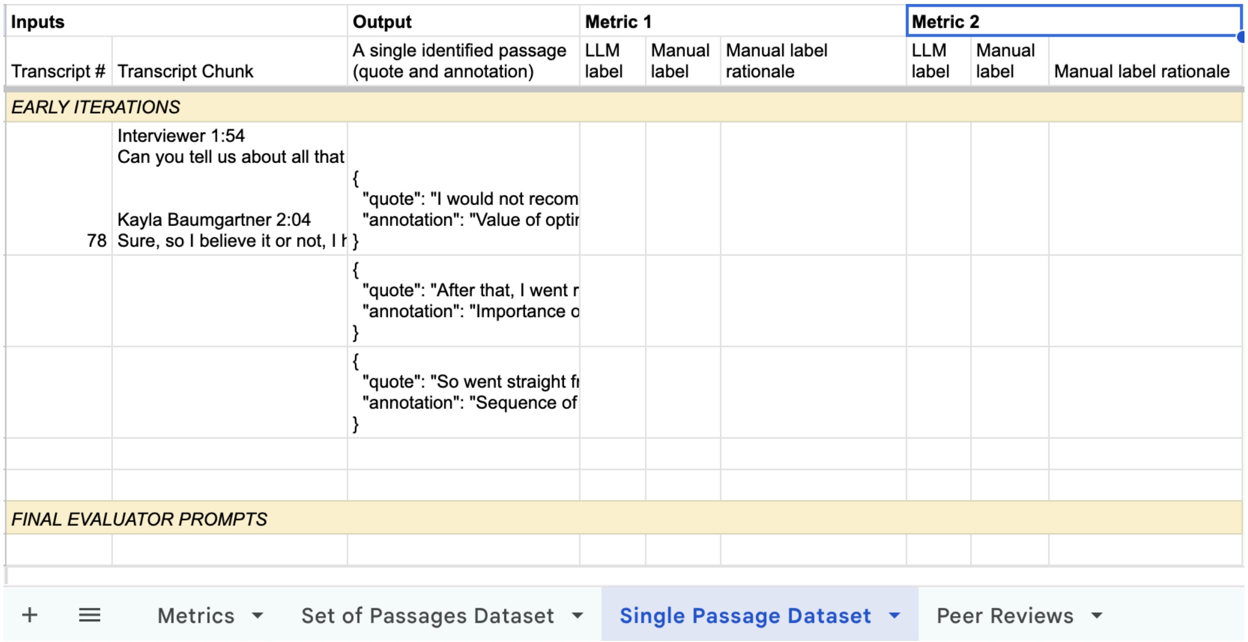
Prepare the dataset curation tabs. The spreadsheet you copied has **two dataset tabs**.

- The 'Set of Passages Dataset' is for evaluating the entire array of identified passages in the generated response output (e.g. whether you missed any, whether there are incorrect things included).
- The 'Single Passage Dataset' is for evaluating a single item within the array of identified passages in the generated response (e.g. the phrasing of the code/theme, the text included in the quote).

Update the columns in these dataset tabs based on your defined evaluation metrics. We have only added two evaluation metrics, but you may add additional columns if you have additional metrics.



Inputs		**Output**	**Rank**	**Metric 1**			**Metric 2**		
Transcript #	Transcript Chunk	Generated passages as an array of objects	Ranking of outputs	LLM label	Manual label	Manual label rationale	LLM label	Manual label	Manual label rationale
EARLY ITERATIONS									
78	Interviewer 1:54 Can you tell us about all that Kayla Baumgartner 2:04 Sure, so I believe it or not, I h	[{ "quote": "I would not rec "annotation": "Value of }, {							
FINAL EVALUATOR PROMPTS									



Inputs		Output	Metric 1			Metric 2		
Transcript #	Transcript Chunk	A single identified passage (quote and annotation)	LLM label	Manual label	Manual label rationale	LLM label	Manual label	Manual label rationale
EARLY ITERATIONS								

78	Interviewer 1:54
	Can you tell us about all that
	Kayla Baumgartner 2:04
	Sure, so I believe it or not, I h
		{
		"quote": "I would not recom
		"annotation": "Value of optir
		}						

		{
		"quote": "After that, I went r
		"annotation": "Importance o
		}						

		{
		"quote": "So went straight fr
		"annotation": "Sequence of
		}						

FINAL EVALUATOR PROMPTS								

Manually evaluate a few generated responses. Pick **five** transcript chunks that are different from each other in the output you expect 078_Kayla Baumgardner Firefighter Paramedic (e.g. one with no relevant passages, one with just one, and one with multiple relevant passages). For these transcript chunks, run your main prompt to generate outputs and manually evaluate the output chunks using your evaluation metrics. You should add your manual evaluations to the 'Set of Passages Dataset' tab and the 'Single Passage Dataset' tab under the "EARLY ITERATIONS" section.

Based on your initial manual evaluation, see if there are other flaws in the generation process that were not captured by your current evaluation metrics. If so, define new evaluation metrics that target the missing aspects to what determines if an output response is high-quality.

Define an evaluator prompt for each evaluation metric. You now need to write a new evaluator prompt for each of the evaluation metrics you defined beforehand. Your evaluator prompt can use the research question, the transcript chunk, and the relevant output (either the array of relevant passages or just a single relevant output).

Your evaluator prompt should also describe the criteria that distinguish a "1" or "0" for that particular evaluation metric. Add your evaluator prompt to the 'Metrics' tab of the spreadsheet.

Run and refine your evaluator prompts. Run your main prompt on the same **5 diverse chunks** and add the output into the 'EARLY ITERATIONS' section of the spreadsheet. Also manually define a 'ground truth' ideal output for each of those 5 diverse chunks (unless the generated output is already a good ground truth. In that case, add a flawed output). Then add the ranking labels for the two outputs for each of your 5 chunks (1 for the ground truth, 2 for the generated / flawed output).

Run your evaluator prompts on the 10 outputs, documenting the results within the ‘EARLY ITERATIONS’ section of your spreadsheet. Remember that some evaluator prompts will take the entire array whereas others will just focus on evaluating a single identified relevant passage within the array:

- For evaluators that evaluate the entire array of identified passages, apply them to the direct prompt output, and record the result (1 or 0) in the ‘SET OF PASSAGES DATASET’ tab,
- For evaluators that evaluate single items within the produced arrays, you should apply them separately to each of the items in the outputs, and record the result (1 or 0) in the SINGLE PASSAGE DATASET’ tab. Since there could be many items across the 10 outputs, you only need to do this for 15 items max.

Review all your LLM-generated evaluation labels to create the manual labels, correcting the LLM labels if there are any you disagree with, along with a short rationale for changing the labeling decision from 0 to 1 or vice versa.

Analyze the differences between the LLM-generated evaluation labels and your manual evaluations. This may include:

- Refining your main prompt after seeing how some generated outputs fail your defined criteria,
- Refining your evaluator prompt based on times when you had to manually intervene to change the LLM-generated label.

Importantly, as you iterate, please 1) update the evaluator prompt in the ‘Metrics’ tab of the spreadsheet, 2) keep adding to your curated dataset.

Apply your final evaluator prompts to ALL data for the given transcript. Once you have finalized your evaluator prompts, apply them to ALL the transcript chunks under the ‘FINAL EVALUATOR PROMPTS’ section for the ‘SET OF PASSAGES DATASET’ tab.

- For the individual assignment, you only need to do this for a single transcript: 078_Kayla Baumgardner Firefighter Paramedic.
- Within the ‘SINGLE PASSAGE DATASET’, you only need to have 30 rows (you do not need to apply it to every single excerpt in the transcript). You can include the 15 rows that you used in your early iterations as part of this 30.

Once you have generated evaluations of each transcript chunk based on all your metrics, record the LLM evaluation labels in the relevant spreadsheet tabs and review them to create manual evaluations with rationales for any incorrect ones.

- You do not need to create ground truths or to do rankings for these transcript chunks

Here are some tips for improving your prompts, given by students who completed this same task:

1. Adding an example in the prompt can drastically improve your results for getting quotes and generating codes in the correct format. Consider adding an example from your manual thematic analysis in Week 1 or from your last iteration in Week 2.
2. When prompting the LLM, consider giving it a role or context to help it understand the task e.g:

    a. "You are a researcher conducting thematic analysis of transcripts….."
    b. "You are a highschool career counsellor/advisor…." (I instructed it to prioritize insights that address common high schooler concerns, such as how to explore different career paths, develop valuable skills, and gain practical experience through internships or volunteering. I emphasized the importance of personal stories, relatable advice, and actionable tips while specifying what to disregard, such as overly technical or irrelevant information. This approach ensured the outputs were engaging and relevant to the target audience.)

3. Be clear with the model about quote lengths or you may end up with very short quotes or entire paragraphs. You can also tell the model to trim quotes to remove unnecessary information from the beginning, end or middle of the quote.

Reflect and submit. Finally, reflect on the individual activity. Submit your responses to the below questions directly on the Google Form along with your **spreadsheet** of evaluations and prompt iterations. **IMPORTANT: MAKE SURE TO GIVE US PERMISSIONS TO ACCESS DOCUMENTS**

1. Please provide a one paragraph reflection of how you defined the evaluation metrics and criteria for your task. What did you learn from your initial manual evaluation of a few transcript chunks on your criteria/metrics?

2. Write a one paragraph reflection on the differences in your manual evaluations and the LLM evaluations. Were there specific metrics on which your manual evaluations differed from the LLM? If yes, why do you think that happened? How did this inform your second iteration of your evaluation prompts?

3. Write a one paragraph reflection on your process of refining your main prompt? What metrics did you particularly have to focus on for improving the quality? What changes did you incorporate in your main prompt to improve those metrics?

Team Assignment Instructions

In-section peer review of evaluation. Among the team members who are present at the section, conduct a peer review of each other's work. We don't have a strong preference for how you assign peer reviewers, but one way to do it is to use a cycle A → B → C → D → A. In other words, if you have 4 team members present (A, B, C, D), have A peer review B, B peer review C, C peer review D, and D peer review A.

Examining your team member's evaluation metrics, evaluation prompts and evaluation labels (in their evaluation spreadsheet), take note of specific ways they could have improved in the metrics defined or prompts. Then:

- Write a two sentence reflection on what your team member did well in. For example, did they define a new metric that helped identify problems in the response output? Were any of their evaluation prompts particularly well-written?
- Write a two sentence reflection on a concrete, specific way in which your partner could have improved their metrics or their evaluation prompt. You should focus on what is most important / would make the biggest impact or improvement.

Note: The course staff will share some initial observations and pointers, but this is only based on a quick review of the individual submissions. You should take what they say into account, but make sure to be detail-oriented in thinking about the submission you are reviewing and how it should improve.

Compile your reviews. Create another copy of the evaluation spreadsheet and compile the reviews from all of the participating team members in the 'Peer Reviews' tab. Specify for each review, who was the reviewer, who was being reviewed, the two sentence reflection on what your team member did well in, and the two sentence reflection on the most important way they could have improved.

We are looking at the quality of your reviews, so the way to maximize your points is to write the most helpful critique that points out the biggest way each team member can improve.

Iterate as a team. Work together to create a single final team version of the evaluation metrics, evaluation prompts and main prompt, building on the work you have already done in the individual submission and considering what you might want to draw from each submission (e.g. specific evaluation metrics, evaluation prompts, techniques for refining the prompts used, etc). Note: you do not have to use something from every individual submission. You should be trying to create a final prompt that produces the highest quality outcomes you can achieve.

In the same copy of the evaluation spreadsheet where you compiled your peer reviews, fill in the 'Metrics' tab with your evaluation metrics and evaluation prompts and then fill out the 'Set of Passages Dataset' tab and the 'Single Passage Dataset' tab by applying your prompts to all **six** of your teams transcripts following the same process as the individual submission to generate outputs for and LLM+manual evaluations. If you identify issues, use these to improve and iterate on your main prompt and evaluator prompts.

- You only need to do this for 8 transcript chunks within each of your 6 transcripts, choosing a diverse set to better stress test and refine your prompts,
- For each of the 8 transcript chunks per transcript, there should be at least two outputs in the 'SET OF PASSAGES DATASET', one that is an ideal / ground truth output and one that is flawed in at least one way. These should be ranked (the ideal / ground truth ranked 1 and the flawed ranked 2). All outputs should have LLM and manual evaluation labels like in the individual assignment,
- For each of the 6 transcripts, there should be at least 15 outputs in the 'SINGLE PASSAGE DATASET'. All outputs should have LLM and manual evaluation labels like in the individual assignment,

Reflect and submit. Submit your responses to the below questions directly on the Google Form along with your spreadsheet of your final prompt and results. **IMPORTANT: MAKE SURE TO GIVE US PERMISSIONS TO ACCESS DOCUMENTS**

1. Write 3-4 sentences on your team's thought process. How did you choose what evaluation metrics to use? What influenced your final prompt changes (main prompt and evaluation prompt)? What further improvements can be made? What feedback did you incorporate (from peer review or your tutor)?
2. Write a sentence or two describing the team dynamics. Were there any challenges you faced working in your team and how did you overcome them?
3. Please list each member of your team, whether they attended and engaged in section discussions, and their specific contributions.

Evaluation Rubrics

Individual submissions will be graded on a Check+, Check, Check-, Minus+, Minus scale according to the below rubric. The purpose of individual submissions is primarily to ensure that all members are contributing to their team, so graders will not be providing feedback on these.

Check+ Outstanding, one of the best in the class (102%),
Check High quality, though not one of the best in minor ways (95%),
Check- Completed the work, but needs significant improvement (80%),
Minus+ Low quality or missing significant portions (40%)
Minus Did not do the work or barely did any work (0%)

Team submissions will be graded according to detailed rubrics to be posted on the rubric spreadsheet