Prompt Engineering for Customer Review Analysis Report

Ngoc Mai Nong

May 2025

1 Methodology

For this assignment, I conducted prompt engineering for customer review analysis using a dataset of Udemy online course customer review. I started with randomly selecting 50 reviews from the original dataset and manually labeled them for each task.

Then, for prompt engineering, I designed three prompts per task to compare their performance across different large language models. In general, the first prompt version is the most direct, consisting of only 1 sentence of instruction (e.g. Classify the sentiment of the following customer review). The second version of the prompt is formulated as a question, followed by an instruction of what the answer should contain (e.g., What is the most significant phrase of praise or complaint from the following review? Extract a key phrase or sentence). The third version assigns a persona to the model as an assistant who is analyzing an online course's customer reviews along with the task at hand (e.g., You are a helpful assistant summarizing customer feedback for an online course. Write a short (1-2 sentence) summary that captures the core experience and opinion of the customer). Additionally, to ensure that I can easily retrieve the response content, I also needed to specify the response format I wanted. Conveniently, the `lm-studio` Python SDK allows me to specify the response structure through the parameter `response_format`.

I selected three models for comparison and evaluation of the prompts: `llama-3.2-3b-instruct`, `gemma-3-4b-it`, and `falcon3-3b-instruct`. These 3 models are relatively comparable in terms of number of parameters (3 to 4 billions) and size that are adequate for the tasks and suitable for the computer resource constraints (my computer has about 8GB of RAM and 4GB of GPU). With each model, I iteratively performed the 3 tasks with the 3 prompt versions and saved the responses for evaluation afterwards.

2 Findings

2.1 Task 1: Sentiment Classification

The macro F1 score for the 3 models across all three prompt versions are reported in Table 1 and the corresponding confusion matrices are displayed in Figure 1. Overall, the gemma model performed the most consistent across all prompt versions and performed the best with prompt version 2. It also made the least number of misclassifications of positive and negative reviews. However, the Falcon model did the best job separating the neutral reviews from the rest, especially with prompt version 3. The Llama model tends to misclassify them as positive reviews whereas the Gemma model tends to classify them as negative.

Latency-wise, the Llama model was the slowest, averaging at about 1.04 to 1.1 seconds per review while the Falcon model was the fastest, at about 0.3 seconds per review.

2.2 Task 2: Key Praise or Complaint Extraction

The average token-level precision, recall, and F1-score for each model with different prompt versions are presented in Table 2. Each model performs best with a different prompt version: The Gemma model performed the best with the first prompt version, the Llama model with the second prompt version, and the Falcon model with the third version.

Falcon is the fastest model with an average processing time per review of 0.4 to 0.5 seconds, while Llama is the slowest, at 1.4 to 1.5 seconds per review.

Table 1: Macro F1 score across different prompt versions and models for task 1

Prompt version	**Macro F1-score**
1. Classify the sentiment of the following customer review as 1 if it is overall positive, 0 if neutral, or -1 if negative.	`llama-3.2-3b-instruct`: 0.72 `gemma-3-4b-it`: 0.70 `falcon3-3b-instruct`: 0.74
2. What is the sentiment of the following customer review? Use the label 1 for positive, 0 for neutral, and -1 for negative.	`llama-3.2-3b-instruct`: 0.46 `gemma-3-4b-it`: 0.83 `falcon3-3b-instruct`: 0.75
3. You are a sentiment analysis assistant. Read the following customer review and classify its overall sentiment as 1 for positive, 0 for neutral, or -1 for negative, based on the customer's tone and opinion.	`llama-3.2-3b-instruct`: 0.67 `gemma-3-4b-it`: 0.72 `falcon3-3b-instruct`: 0.64

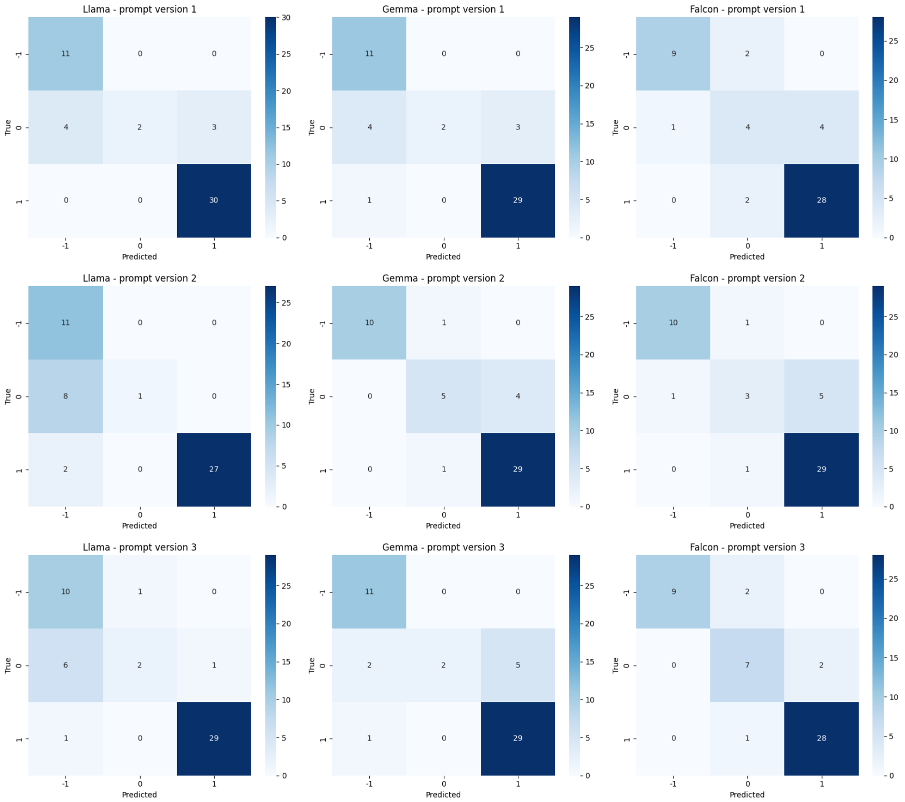

Figure 1: Confusion matrices across different prompt versions and models for task 1

2.3 Task 3: Review Summarization

The average ROUGE-1, ROUGE-L, and METEOR scores for each model across different prompt versions are reported in Table 3. Overall, the most simple prompt version - version 1, yielded the best results across all three models, whereas the more "engineered" prompt versions ended up with lower ROUGE-L and METEOR scores. With prompt version 1, the three models produced similar ROUGE-L scores, which measure the longest common subsequence between generated and reference summaries. However, the Llama and Gemma models achieved higher METEOR scores, which account for synonyms, and higher ROUGE-1 scores, reflecting unigram overlap, compared to the Falcon model.

Nevertheless, the Falcon model outperforms both the Llama and Gemma models in terms of latency, processing a review in just 0.82 seconds, compared to 3.5 seconds for Llama and 2.3 seconds for Gemma.

Table 2: Average token-level Precision / Recall / F1 across different prompt versions and models for task 2

Prompt version	**Average Token-level Precision / Recall / F1**
1. Extract the most significant phrase of praise or complaint from the following review.	`llama-3.2-3b-instruct`: 0.37 / 0.44 / 0.37 `gemma-3-4b-it`: 0.52 / 0.40 / 0.41 `falcon3-3b-instruct`: 0.39 / 0.43 / 0.38
2. What is the most significant phrase of praise or complaint from the following review? Extract a key phrase or sentence.	`llama-3.2-3b-instruct`: 0.41 / 0.45 / 0.40 `gemma-3-4b-it`: 0.39 / 0.38 / 0.33 `falcon3-3b-instruct`: 0.32 / 0.49 / 0.35
3. You are an assistant analyzing customer feedback on online courses. Your task is to extract a key phrase of praises or complaints mentioned in the review.	`llama-3.2-3b-instruct`: 0.26 / 0.23 / 0.23 `gemma-3-4b-it`: 0.32 / 0.30 / 0.28 `falcon3-3b-instruct`: 0.44 / 0.44 / 0.40

Table 3: Average ROUGE-1 / ROUGE-L / METEOR across different prompt versions and models for task 3

Prompt version	**Average ROUGE-1 / ROUGE-L / METEOR**
1. Provide a 1-2 sentence summary of the given review.	`llama-3.2-3b-instruct`: 0.44 / 0.33 / 0.36 `gemma-3-4b-it`: 0.45 / 0.34 / 0.37 `falcon3-3b-instruct`: 0.41 / 0.32 / 0.30
2. What is the overall experience and opinion of the customer based on the following review? Summarize it in 1-2 sentences.	`llama-3.2-3b-instruct`: 0.39 / 0.28 / 0.32 `gemma-3-4b-it`: 0.40 / 0.30 / 0.31 `falcon3-3b-instruct`: 0.39 / 0.29 / 0.27
3. You are a helpful assistant summarizing customer feedback for an online course. Write a short (1-2 sentence) summary that captures the core experience and opinion of the customer.	`llama-3.2-3b-instruct`: 0.40 / 0.29 / 0.36 `gemma-3-4b-it`: 0.39 / 0.30 / 0.28 `falcon3-3b-instruct`: 0.36 / 0.28 / 0.28

3 Final Leaderboard Table

Table 4: Leaderboard: Prompt and Model Rankings by Task

Task	**Prompt**	**Prompt Rank**	**Model Ranking**		
			1st Place	**2nd Place**	**3rd Place**
1	1	1	`falcon3-3b-instruct`	`llama-3.2-3b-instruct`	`gemma-3-4b-it`
	2	2	`gemma-3-4b-it`	`falcon3-3b-instruct`	`llama-3.2-3b-instruct`
	3	3	`gemma-3-4b-it`	`llama-3.2-3b-instruct`	`falcon3-3b-instruct`
2	1	1	`gemma-3-4b-it`	`falcon3-3b-instruct`	`llama-3.2-3b-instruct`
	2	2	`llama-3.2-3b-instruct`	`falcon3-3b-instruct`	`gemma-3-4b-it`
	3	3	`falcon3-3b-instruct`	`gemma-3-4b-it`	`llama-3.2-3b-instruct`
3	1	1	`gemma-3-4b-it`	`llama-3.2-3b-instruct`	`falcon3-3b-instruct`
	2	3	`gemma-3-4b-it`	`llama-3.2-3b-instruct`	`falcon3-3b-instruct`
	3	2	`llama-3.2-3b-instruct`	`gemma-3-4b-it`	`falcon3-3b-instruct`