CEU Review Analysis

I. Introduction

In this study, we evaluate the performance of the LLaMA 3 8B language model on customer review analysis tasks. We focus on three specific tasks: (A) sentiment classification labeling each review as Positive, Neutral, or Negative, (B) extracting the key complaint or praise from each review, and (C) summarizing each review in 1–2 sentences. For each task, we designed three different prompt variants (A, B, C) to explore prompt engineering strategies and their impact on the model's outputs. We run the model on a test set of 20 TripAdvisor reviews of a gelato shop in Rome, for which human-annotated true sentiment and true key phrase are provided. We then assess model performance using metrics for each task and compare the prompt variants to determine which elicited the best results from LLaMA 3. Finally, we discuss the results, compare them to expectations from larger or fine-tuned models, and reflect on the viability of using LLaMA 3 8B for these review analysis tasks.

II. Prompt Design for Each Task

We created three prompt versions for each task, varying in the level of instruction detail and format, to see how they influence LLaMA's outputs:

Task A (Sentiment Classification):
Prompt A: A minimal directive – e.g., *"Classify the sentiment of this review as Positive, Neutral, or Negative:"* followed directly by the review text. This straightforward prompt relies on the model to complete the task with no additional guidance.
Prompt B: A role-oriented prompt – e.g., *"You are a sentiment classifier… Output one of the following: Positive, Neutral, or Negative. Content: [Review]."* This prompt explicitly tells the model its role and the expected output format (one of the three labels). The aim is to focus the model on producing just the category name.
Prompt C: A detailed instruction – e.g., *"Sentiment Analysis Task: Please assess the sentiment… take everything into account (every feeling, sarcasm etc.). Choose from: Positive / Neutral / Negative. Content: '[Review]'."* This version gives the most explicit guidance, including mention of considering sarcasm/negation, hoping to improve handling of nuanced language. It sets the stage as a formal task to encourage a concise label output.

Task B (Key Complaint/Praise Extraction):
Prompt A: A direct instruction – e.g., *"Extract the most important sentence or phrase that expresses a key complaint or praise… Return no more than 4 words. Do not explain."* followed by *"Review: [text]"*. This prompt straightforwardly asks for the key phrase without any rationale or examples, and explicitly limits the answer length to encourage brevity.
Prompt B: A role and rule-based prompt – e.g., *"You are a customer sentiment analyst… extract the core complaint or praise, no more than 4 words… Extract only the most significant phrase… Do not write explanations. Review: [text]"*. This provides some context ("customer sentiment analyst") and reiterates the 4-word limit and requirement to pick the **core** issue or praise. The intention is to guide the model to focus and not deviate (similar to Prompt A but with slightly more clarity and authority).

Prompt C: A structured prompt with examples – it explicitly states the task and provides bullet-point instructions on what to output depending on sentiment:
*Task: Identify the key complaint or praise… (instructions)… Review: "[text]" … Output: *
This version is the most elaborate; it outlines how to decide what to extract (depending on if the review is positive or negative) and emphasizes the 4-word limit.

Task C (Summarization):
Prompt A: A simple instruction – *"Summarize this customer review in 1-2 sentences:"* followed by the review text. This is a basic zero-shot summary request, leaving it entirely to the model how to condense the content.
Prompt B: A role prompt with constraints – e.g., *"You are a summarization assistant… return a concise summary. It should be 1–2 sentences and capture the main idea and sentiment… factual and in third-person."* followed by the review. This guides the model on style (third-person, no added opinions) and content (main idea and sentiment). The goal is to reduce rambling or inclusion of irrelevant details by clearly stating the requirements.
Prompt C: A few-shot prompt – we prepend two example reviews and their correct 1-2 sentence summaries, then ask *"Now, summarize the following review…"*. The examples illustrate how a positive review's summary should look (enthusiastic, highlighting what was loved) and how a negative review's summary should look (highlighting dissatisfaction).

III. Evaluation Methodology

We evaluated each model output against human-provided labels (for tasks A and B) or reference outputs (for task C) using appropriate metrics:
Sentiment Classification (Task A): We computed accuracy (overall percentage of correct Positive/Neutral/Negative predictions) and macro-averaged F1 score. We also generated a confusion matrix.
Key Phrase Extraction (Task B): We evaluated two aspects: **Exact Match (EM)** accuracy and **token-level Precision/Recall/F1**. EM measures the percentage of reviews where the model's extracted phrase exactly matches the gold-standard phrase provided by humans. This is a strict metric. We also calculated precision and recall on the set of words in the predicted phrase versus the set of words in the true phrase, then their F1. This token-level F1 (macro-averaged across all reviews) gives credit for capturing some of the correct keywords even if not an exact string match and is more forgiving when the model's phrasing differs slightly from the ground truth. Precision tells us how much extra, unrelated words the model included, and recall tells us how much of the true key idea was captured.
Summarization (Task C): We used a **model-based reference**: we took high-quality summaries generated by GPT-4 as a proxy for ground truth. We then computed **ROUGE-1** and **ROUGE-L** scores to compare LLaMA's summaries with these reference summaries. ROUGE-1 essentially measures the overlap individual words between the model's summary and the reference. ROUGE-L measures overlap based on the longest common subsequence. We report the F1 form of these ROUGE scores (harmonic mean of precision and recall). A higher ROUGE indicates that the summary captured more of the same information as the reference (note that this doesn't perfectly equate to quality, but it's a useful automated comparison).

All evaluations were done on the same set of 20 reviews to ensure comparability. We did not fine-tune the model on any task; these are zero-shot or few-shot prompt-based results. All analysis code and metrics were computed in Python.

IV. Results and Analysis

Task A. Across the three prompting strategies, LLaMA 3 8B demonstrated strong performance on positive and negative reviews but consistently struggled with the neutral class. With the simplest prompt (A), the model reached 75% accuracy and a macro F1 of 0.51, correctly identifying most positives but mislabeling both neutral examples and a few positive reviews. Introducing a role-based prompt (B) raised accuracy to roughly 85% and F1 to 0.58, reducing some errors on positive reviews yet still failing to reliably catch neutrality. The most detailed prompt (C) achieved the highest accuracy (90%) and F1 (≈ 0.59), correctly classifying nearly all positive and negative cases, but it still labeled every neutral review as polar sentiment. In other words, despite increasingly specific instructions, the model defaulted to assigning positive or negative labels. GPT-4, on the other hand, achieved a lower overall accuracy of 80% with the same prompt, but it more evenly distributed predictions across all three classes, producing a macro F1 of 0.57. This indicates better handling of class imbalance and nuanced sentiment.

Prompt	Sentiment Accuracy	Sentiment F1 (macro)	Extraction Exact Match	Extraction Token F1	Summarization ROUGE-1	Summarization ROUGE-L
A (Basic)	0.75 (75%)	0.51	0.0 (0%)	0.25	0.45	0.40
B (Role/Format)	0.85 (85%)	0.58	0.05 (5%)	0.30	0.50	0.45
C (Detailed/Few-shot)	0.90 (90%)	0.59	0.05 (5%)	0.35	0.55	0.50

Table 1: Performance of LLaMA 3 8B under each prompt version. Prompt C is the top-performer in most cases (highest values in each column are bolded).

Prompt	Sentiment Accuracy	Sentiment F1 (macro)	Extraction Exact Match	Extraction Token F1
A (Basic)	0.75 (75%)	0.51	0.00 (0%)	0.28
B (Role/Format)	0.75 (75%)	0.52	0.05 (5%)	0.35
C (Detailed/Few-shot)	**0.80 (80%)**	**0.57**	**0.05 (5%)**	**0.39**

Table 2: Performance of GPT-4 under each prompt version. Prompt C is the top-performer in most cases (highest values in each column are bolded).

Task B. LLaMA 3 8B struggled significantly with the key complaint/praise extraction task across all three prompts: exact-match accuracy remained near zero (0–5%), reflecting the difficulty of verbatim extraction, and even the most permissive structured prompt (C) only occasionally hit the gold phrase by chance. Token-level F1 scores illustrate the same challenge—rising modestly from roughly 0.25 under the basic prompt to about 0.35 with the detailed, rule-based prompt—meaning the model captured only a fraction of the correct keywords. While Prompt B's emphasis on "core complaint or praise" improved precision by encouraging concise, relevant phrases (e.g. "amazing choice" vs. "amazing choice of flavours"), and Prompt C further boosted recall by enforcing output format and sentiment-dependent guidance ("Authentic gelato, generous portions"), overall overlap remained low. These results demonstrate that without fine-tuning or task-specific training, LLaMA 3 8B's ability to zero-shot extract the single most significant four-word snippet is quite limited, even when heavily constrained by prompt engineering. GPT-4 showed notably stronger performance on the key complaint/praise extraction task, although both models struggled with exact matches due to the inherently subjective and open-ended nature of the task.

Taks C. In summarization all three prompts generated coherent outputs but differed in content richness and alignment with GPT-4's reference summaries, as measured by ROUGE scores. Prompt A produced minimal, often generic summaries with low ROUGE and ROUGE-L, missing key details. Prompt B improved on this by providing more specific and factual content, resulting in ROUGE-1 ~0.50 and ROUGE-L ~0.45. Prompt C performed best, leveraging few-shot examples to guide the model toward richer, more structured summaries, with ROUGE-1 ~0.55 and ROUGE-L ~0.50. While LLaMA 3 8B under Prompt C still fell short of the detail and accuracy that fine-tuned models or GPT-4 can deliver, it consistently captured more relevant content.

V. Conclusion and Recommendations

Cost. When comparing costs between GPT-4 and LLaMA 3 8B, the difference lies primarily in usage vs. infrastructure. GPT-4 incurs a small per-review charge via its API—about $0.003 for classification/extraction and $0.004 for summarization, making the total cost for 227 reviews under $1. However, this scales linearly, and analyzing tens of thousands of reviews (e.g., 10,000 summaries for ~$40) can accumulate over time, especially for regular or large-scale tasks. In contrast, LLaMA 3 8B, while free to use as an open-source model, requires hardware or cloud infrastructure. For small datasets, the cost is minimal, even using a cloud GPU for processing 227 reviews may cost under $1. The key advantage of LLaMA 3 8B is the absence of per-call fees, making it far more cost-efficient for recurring or large-scale use cases, assuming you have the expertise and infrastructure to manage the setup. Thus, GPT-4 is ideal for one-off or small tasks with no setup burden, while LLaMA 3 8B becomes financially preferable for ongoing high-volume analysis.

Latency. Both perform well for small-scale or interactive applications, but differ in speed and scalability. GPT-4, accessed via API, typically returns classification or extraction responses in about 1 second, and summaries in 1.5–3 seconds, with rate limits restricting large-batch throughput unless carefully parallelized. In contrast, LLaMA 3 8B, when run locally on a decent GPU, delivers sub-second latency for short tasks—summaries in 0.5–1.5 seconds and classifications nearly instantaneously. While CPU inference is slower, the key advantage of LLaMA 3 8B lies in its flexibility: with no API rate limits, it can process large batches more

efficiently by leveraging hardware scaling. For real-time or high-throughput scenarios, the open model can outperform GPT-4, assuming sufficient compute. However, GPT-4 offers more consistent latency, while LLaMA's performance may vary slightly based on system load—something manageable with good engineering.

GPT-4 vs Open Model: GPT-4 come out as the strongest choice for analyzing Tripadvisor reviews of Giolitti across sentiment classification, key insight extraction, and summarization. It provides high accuracy and rich, fluent summaries. The open LLaMA-3 8B model, while considerably improved through prompt engineering, still lags in capturing nuances and detail.

When to use which:
If **quality and accuracy** are top priority, GPT-4 is worth the cost. Its outputs would likely require minimal oversight.
If **cost or data privacy** is critical, an open model like LLaMA can be used with the understanding that its insights might need review. One could also consider a larger open model (like LLaMA-2 13B or 70B) which might narrow the quality gap, though those come with inference speed and hardware costs.
For **real-time analysis** (lots of queries per second), an open model might be more scalable. But for periodic batch reports (say a weekly analysis of new reviews), GPT-4's cost is quite manageable and the ease of not maintaining your own model server is a plus.

Future Improvements: For the open model, one could explore fine-tuning it on a small labeled dataset of reviews to improve sentiment and extraction specifically. That might boost its accuracy significantly. We could also ensemble the models – e.g., use GPT-4 to label a large dataset, fine-tune LLaMA on it, potentially getting a cheaper model with closer performance. Another approach is using GPT-4 outputs as a reference to evaluate or correct the open model outputs (GPT-4 as an editor). Those are beyond our scope but worth mentioning.

We conclude that GPT-4 with refined prompts is the recommended solution for analyzing customer reviews, other than sentiment analysis, when quality is needed. The open-source model is a viable option for easy tasks as sentiment analysis or for those with budget constraints or needing on-premises deployment.