Structured Insight Generation from Mental Health Counseling Reviews: A GPT Prompt Engineering Enhanced Approach

Sanchari Biswas[1], **Hassan A. Shafei**[2], **Chiu C. Tan**[1]

[1]Department of Computer and Information Sciences, Temple University
sanchari.biswas@temple.edu, chiu.tan@temple.edu

[2]College of Engineering and Computer Sciences, Jazan University, Jazan, Saudi Arabia
hshafei@jazanu.edu.sa

Abstract

This paper explores various prompting techniques and database structures in the context of GPT-based models to extract and evaluate insights from mental health counseling reviews. We develop a pipeline that processes raw data, applies different configurations of GPT prompting (including different batch sizes, with and without examples and explanations), and compares GPT-derived ratings with a manually established ground truth. Through systematic experimentation, we assess the accuracy of each prompting approach, and our findings highlight the impact of data processing and batch prompting configurations on model accuracy. Our results provide guidance for leveraging large language models in structured, sensitive data contexts such as mental health review analysis.

Introduction

The widespread adoption of online mental health counseling has led to an increasing volume of publicly available feedback, such as reviews on platforms like Google. These reviews provide valuable insights into user experiences, satisfaction, and the perceived quality of care offered by mental health services. However, given the unstructured nature of this feedback, extracting structured, actionable insights from these reviews is challenging. Manually analyzing such a large dataset for attributes like Friendliness, General Rating, Flexibility, Ease, and Affordability is labor-intensive, prone to bias, and lacks scalability. As such, there is growing interest in employing natural language processing (NLP) techniques, specifically large language models like GPT (Generative Pre-trained Transformer), to interpret and summarize these reviews.

Recent advancements in generative models, particularly the GPT family, have shown promising capabilities in various NLP tasks, including summarization, sentiment analysis, and question answering. With refined prompt engineering, GPT-based models can generate structured insights from unstructured text data, making them valuable tools for domains where data exists in text-heavy formats. In the context of mental health counseling reviews, where subtle nuances and complex themes are common, prompt engineering enables these models to capture specific aspects of user feedback accurately. However, significant challenges remain in ensuring that the insights generated are both accurate and reliable, as GPT's responses can vary widely depending on how it is prompted.

Prompt engineering has emerged as a critical technique to improve the performance and reliability of large language models. Prompting refers to the strategic formulation of input questions or instructions to guide models in generating accurate and contextually relevant responses. Studies have demonstrated that different prompt configurations—such as phrasing, context length, or using examples—can significantly influence the model's output (Brown 2020). In sensitive domains like mental health, where accuracy is paramount, the choice of prompt design can impact the quality and interpretability of insights derived from user-generated data. Thus, this project explores various prompting strategies, including single vs. batch prompting, and methods that involve additional contextual examples or explanations, to enhance the model's ability to generate accurate and interpretable insights.

Additionally, data preprocessing plays a crucial role in determining the quality of insights derived from language models. Raw review data, especially from open-source platforms like Google, often contains noise, spelling variations, and informal language, which may hinder the model's ability to consistently interpret the data. Cleaning and processing the data before feeding it into GPT can potentially improve model performance by removing irrelevant or misleading information. This study assesses the impact of data preprocessing by comparing the model's performance on raw versus cleaned review data across various prompting techniques.

The project aims to determine the optimal prompting technique and data configuration to achieve high alignment with a manually established ground truth. By establishing a ground truth, this study provides a baseline against which different prompting and data processing strategies can be measured. We systematically compare single and batch prompting, as well as batch prompting with examples and explanations, for both raw and cleaned data. Specifically, the project evaluates the accuracy of these strategies by comparing GPT's ratings on six attributes—Ranking, Friendliness, General Rating, Flexibility, Ease, and Affordability—with human-assigned scores.

Our findings contribute to the broader field of AI-driven mental health analysis by identifying prompt engineering techniques that enhance the interpretability and accuracy of language models in structured review analysis. This study also sheds light on the role of data preprocessing and the impact of structured versus unstructured data formats on model performance. Ultimately, the goal is to create a methodological framework for applying GPT-based models to mental health review analysis in a way that maximizes interpretability, accuracy, and reliability.

To answer whether GPT models, when effectively prompted, produce structured, attribute-based insights that align closely with human-derived ground truths, we address the following research questions:

1. How does prompt configuration (e.g., single versus batch, optimum batch size, with or without examples and explanations) affect the accuracy of GPT-generated ratings for mental health counseling reviews?
2. What is the impact of data preprocessing on the model's performance, and does cleaned data consistently lead to more accurate and reliable insights than raw data?

This paper will discuss the methodology and results in detail, demonstrating how prompt engineering and data handling can refine GPT's analysis capabilities in the context of mental health counseling reviews. The results are expected to provide actionable insights into the best practices for leveraging large language models in fields where interpretive accuracy is essential.

Related Work

The use of large language models (LLMs), particularly generative models like GPT-3 and GPT-4, has significantly advanced natural language processing applications in interpreting and analyzing unstructured data across a wide range of domains, including sentiment analysis (Liu 2017), opinion mining (Cambria et al. 2020), and health informatics (Shickel et al. 2018). These studies highlight the potential for LLMs to automate qualitative analysis in sensitive domains; however, the accuracy and reliability of these insights depend heavily on prompt engineering, data structure, and model configuration.

Prompt engineering, the process of crafting inputs to guide model outputs, is essential for improving model accuracy and relevance. As noted by (Brown 2020), prompt structure—including phrasing, examples, and contextual length—significantly influences model output, impacting the consistency and reliability of responses. This impact is especially notable in healthcare contexts where interpretative precision is critical. Our study builds on this foundation by experimenting with various prompt configurations, such as single, batch, and example-enhanced prompts, to determine the optimal approach for structured analysis of mental health reviews.

Additionally, the evolving capabilities of LLMs in knowledge retention and retrieval present both opportunities and challenges in handling complex, knowledge-intensive tasks. Large pre-trained models like GPT-3 and GPT-4 store extensive factual knowledge within their parameters (Min et al. 2023; Li et al. 2024). However, their ability to retrieve and manipulate stored information accurately remains limited, as demonstrated in tasks requiring precise information retrieval and provenance tracking (Lewis et al. 2020; Petroni et al. 2020). This limitation has motivated the development of Retrieval-Augmented Generation (RAG) frameworks, which combine parametric memory (from language models) with non-parametric memory (retrieved knowledge) to improve factual accuracy and context relevance (Lewis et al. 2020). The RAG model architecture has demonstrated particular promise in healthcare by supporting tasks requiring reliable information retrieval, offering a mechanism for LLMs to access and integrate external structured data dynamically.

In line with the RAG approach, recent advancements in database-interactive NLP systems have shown that combining LLMs with retrieval mechanisms allows for enhanced information synthesis in both structured and semi-structured contexts (Karpukhin et al. 2020). By enabling dynamic querying of databases, RAG models facilitate more precise question-answering in knowledge-dense fields, allowing for accurate, contextually aware responses. This hybrid approach has been successfully applied in service review analysis, where LLMs can retrieve and synthesize structured information for contextually relevant output. For example, LLM-based database interaction frameworks have demonstrated improved performance in healthcare-specific question-answering, where LLMs are augmented with relational or NoSQL databases to generate responses that align with real-world clinical contexts (Karpukhin et al. 2020).

Our project leverages these insights by implementing a retrieval-augmented strategy that integrates GPT-based models with multiple prompting strategies. This setup, akin to the RAG framework, allows our model to retrieve structured and semi-structured data dynamically, thereby enhancing accuracy and relevance in analyzing mental health reviews.

Insights using GPT

Figure 1 illustrates the flow of our work. Our objective is to test different prompting strategies to figure out which strategy returns the most accurate responses. The figure outlines the workflow, starting from data collection and preprocessing, to the evaluation of GPT outputs using varied prompting strategies. This workflow highlights the step-by-step methodology we employed to assess the impact of prompts on data analysis accuracy and consistency.

We collect the data from publicly available Google reviews, then we clean this data. We next evaluate ground truth manually for several attributes, which we discuss in details in the following sections. We then subject both our raw data and cleaned data to several prompting strategies, calculating the accuracy for each, concluding which prompting strategy performs the best in our scenario.

Data set generation

We begin by collecting raw data, consisting of comments from Google reviews on five mental health centers in Birm-

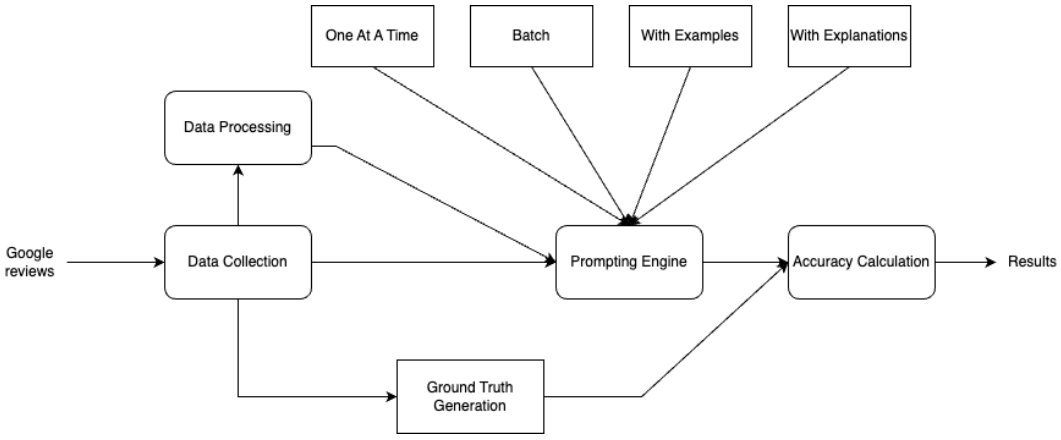

Figure 1: Flow Diagram

ingham, Alabama. The data contains the name of the commenter, the year the comment was posted, the rating out of 5, and the body of the comment. This data is stored in separate CSV files, one file for each health center.

Next, we establish our ground truth. We manually evaluate the comments, giving each a score from 0 to 5 on Ranking, Friendliness, General Rating, Flexibility, Ease, and Affordability, 1 being lowest and 5 being highest. We give a 0 when the attribute cannot be evaluated from the comment.

Next, we clean and preprocess the dataset of mental health counseling reviews for subsequent analysis. We perform text normalization tasks, such as removing special characters, fixing spelling errors, and standardizing formatting. The cleaned data is stored in a folder for downstream processing, enabling comparative analyses of model performance on unprocessed vs. processed data.

Prompting strategies

We consider the following prompting strategies.

1. **Batch Prompting:** After preprocessing, both raw and processed comments are grouped into batches of sizes 1, 10, 25, 50, and 100 and are prompted consecutively. The resultant scores are collected and saved, which are then aggregated to get the overall average scores for each prompting type.
2. **Prompt With Examples:** Prompts include three annotated examples, providing the model with context on how to rate comments. Similar to the previous scenario, both raw and processed comments are prompted consecutively. The resultant scores are collected and saved, which are then aggregated to get the overall average scores for each prompting type. The table 1 is an example provided along with the prompt for this strategy.
3. **Prompt With Explanations:** The model is instructed to provide an explanation for its ratings, adding interpretive value to the outputs. Again similar to the previous scenarios, both raw and processed comments are prompted consecutively and the resultant scores collected, saved, and aggregated to get the overall average scores for each prompting type. The following is an example provided along with the prompt for this strategy.

   The table 2 is an explanation returned by GPT along with the response for this strategy.

Exploring these prompting strategies allows us to systematically evaluate the performance of GPT in diverse real-world scenarios and assess how prompt design impacts the quality of model outputs. The "One At A Time" and "Batch Prompting" approaches provide a baseline to measure computational efficiency and consistency, while advanced techniques like "Prompting with Examples" and "Prompting with Explanations" add layers of contextual understanding and interpretability. Including annotated examples in the prompt has been shown to guide the model in aligning outputs more closely with the desired structure and scoring criteria, as demonstrated by (Brown 2020)." Similarly, prompting with explanations enhances trustworthiness and transparency in model decisions, aligning with insights from (Ribeiro, Singh, and Guestrin 2016).

Evaluation Results

Ground Truth

To establish a reliable ground truth, we manually reviewed and validated the collected data from five counseling centers in Birmingham, Alabama. This ground truth dataset was used to evaluate the accuracy of scores assigned to user comments by the LLM for different loading prompts. The accuracy represents how closely the LLM-generated aggregate scores match the manually established ground truth.

The accuracy for the prompts is computed using the following formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

Batch Sizes

To evaluate the effect of batch sizes, we experimented with batches of 1, 10, 25, 50, and 100 comments. Our findings indicate that smaller batch sizes, particularly those below 10,

LLM Prompt with Example	Explanation
Alabama Psychiatry and Counseling, Mellanie Herard, 4, 2024, love Dr. Whitt. 5 stars rating solely based on appointment having future appointments Dr. Rabbani. For the above comment, we can give it a Ranking of 5, Friendliness of 4, Rating of 4, Flexibility of 4, Ease of 4, Affordability of 0, since the comment doesn't mention anything about affordability.	We provide an example comment and tell the LLM how to score it.

Table 1: Example Prompt Strategy

LLM Response with Explanation	Explanation
Red Peter Panda, 2.0,The client had technical issues with the online portal system., 2.0,The client faced issues with the portal system., 0.0,No explanation provided, 2.0,The client had issues with scheduling and canceling appointments., 2.0,The client faced difficulties using the online portal system., 2.0,No mention of affordability.	When we ask the LLM to provide us with an explanation as well of its scoring, it provides us details of its scoring reasons.

Table 2: Explanation Prompt Strategy

yield lower accuracy. This decrease in accuracy may be attributed to the limited context provided by smaller batches, as some comments are very short and lack sufficient information to derive accurate scores. Additionally, the computational overhead grows linearly with batch size, as the time required for processing each comment averages 0.5 seconds.

Figure 2 illustrates the accuracy for different batch sizes, demonstrating that larger batch sizes generally perform better. However, the trade-off between accuracy and processing time should be considered, especially for real-time applications. Larger batch sizes might also introduce more noise and could make the model more biased and prone to hallucination.

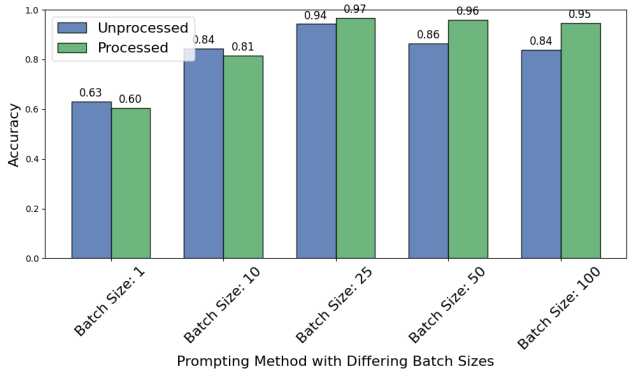

Figure 2: Accuracy Plot for Prompts with Different Batch Sizes

Prompting Strategies

We extended our evaluation to assess two advanced prompting strategies:

1. Providing the LLM with a few examples of how comments should be rated:
   (a) Raw comments with examples
   (b) Processed comments with examples
2. Requesting the LLM to provide explanations for the assigned ratings:
   (a) Raw comments with explanations
   (b) Processed comments with explanations

The results are presented in Figure 3. We observed that including examples in the prompts improved accuracy, as it provided the LLM with a clearer understanding of the expected output. Accuracy further increased when explanations were requested likely because providing explanations forces the model to consider its reasoning process more carefully. This extra layer of reasoning may help the model make more thoughtful and precise predictions, especially when the scoring criteria are nuanced or context-dependent.

Interestingly, the batch prompting strategy with explanations for raw, unprocessed data performed better than for processed data. We speculate that this could be due to sentiment analysis being more effectively conducted on unprocessed data, which retains its original emotional and contextual nuances. However, this remains a hypothesis, as our work is preliminary. A larger dataset is required for a more robust evaluation of these observations.

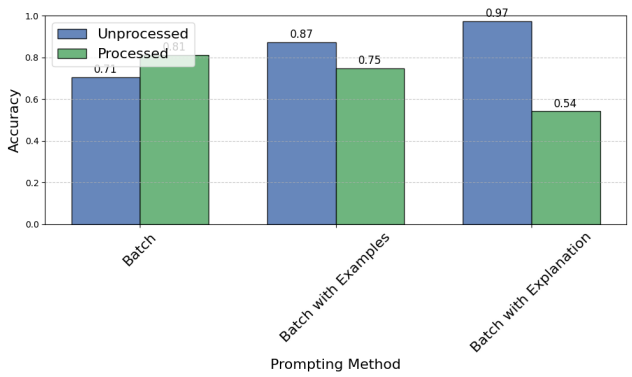

Figure 3: Accuracy Plot for Types of Batch Prompting

Query Runtime Performance

To further analyze the efficiency of different prompting strategies, we measured the average response time per query in multiple configurations. The runtime is influenced by factors such as batch size, the complexity of prompts (e.g., examples or explanations), and the inherent processing speed of the model.

Tables 3 and 4 present the average query response time for each prompting strategy, measured over the five datasets. The results indicate that batch processing significantly reduces per-query latency, with diminishing returns as batch size increases. Notably, batch sizes of 25 and 50 achieve optimal balance between efficiency and response time stability. Additionally, prompting with examples slightly increases the response time due to additional input context, while prompting with explanations incurs the highest computational cost due to the model's extended reasoning and text generation requirements.

These findings suggest that selecting an appropriate batch size and prompt type is crucial for optimizing response time while maintaining accuracy in structured insight generation from mental health reviews.

Prompting Strategy	**Avg. Query Time (sec)**
Single Prompting	0.56
Batch Prompting (10)	0.56
Batch Prompting (25)	0.49
Batch Prompting (50)	0.50
Batch Prompting (100)	0.50
Prompting with Examples	0.51
Prompting with Explanations	0.90

Table 3: Unprocessed Data Runtime Analysis

Limitations

Our evaluation has several limitations. First, the dataset used for testing is relatively small, which may limit the generalization of our results. To draw more definitive conclusions, future work should include larger and more diverse datasets.

Prompting Strategy	**Avg. Query Time (sec)**
Single Prompting	0.49
Batch Prompting (10)	0.49
Batch Prompting (25)	0.53
Batch Prompting (50)	0.45
Batch Prompting (100)	0.41
Prompting with Examples	0.50
Prompting with Explanations	0.72

Table 4: Processed Data Runtime Analysis

Second, our ground truth scoring is based on manual evaluation, which introduces the possibility of human bias. While efforts were made to ensure consistency, subjective interpretations could influence the ground truth scores, impacting the overall evaluation.

Finally, as this is preliminary work, some of our speculations, such as the improved performance of batch prompts with explanations for raw data, require further validation with larger datasets and more rigorous testing. These limitations highlight the need for additional research to build on our findings and address these gaps.

Conclusion

This study investigates the application of large language models (LLMs), specifically GPT-based models, for extracting structured insights from mental health counseling reviews. Through a comprehensive evaluation of various prompting strategies, data preprocessing techniques, and retrieval-augmented generation (RAG) mechanisms, we demonstrated significant findings that advance the understanding of LLM performance in this domain.

The results reveal that data preprocessing and batch prompting consistently improve the accuracy of LLM-generated scores when compared to individual processing of raw comments. Advanced prompting strategies, such as providing examples and asking for explanations, further enhance accuracy, illustrating the value of contextual guidance in aligning LLM outputs with human-generated ground truth. However, the trade-off between accuracy and interpretability is evident in the reduced performance observed when LLMs are required to generate explanations on processed data. This highlights the need to carefully balance interpretability and precision in real-world applications of generative models.

In conclusion, our work provides actionable insights into how prompt engineering, preprocessing, and retrieval-augmented strategies can optimize the use of LLMs for analyzing mental health reviews. These findings contribute to the growing body of knowledge on applying LLMs in sensitive, insight-driven domains like mental health. Future work could explore further refinements in prompting strategies, as well as the integration of domain-specific knowledge bases, to enhance both accuracy and interpretability. This research provides a foundational framework for deploying AI-driven decision-support tools in healthcare service evaluation, ensuring both reliability and transparency.

References

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cambria, E.; Li, Y.; Xing, F. Z.; Poria, S.; and Kwok, K. 2020. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 105–114.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, J.; Tang, T.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9): 1–39.

Liu, B. 2017. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2): 1–40.

Petroni, F.; Piktus, A.; Fan, A.; Lewis, P.; Yazdani, M.; De Cao, N.; Thorne, J.; Jernite, Y.; Karpukhin, V.; Maillard, J.; et al. 2020. KILT: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Shickel, B.; Loftus, T. J.; Ozrazgat-Baslanti, T.; Ebadi, A.; Bihorac, A.; and Rashidi, P. 2018. DeepSOFA: a real-time continuous acuity score framework using deep learning. *ArXiv e-prints*, 1802.