Decoding Sentiment with Large Language Models

Comparing Prompting Strategies Across Hard, Soft, and Subjective Label Scenarios

Timur Oberhuber[1]

Supervisor(s): Luciano Cavalcante Siebert[1]**, Amir Homayounirad**[1]**, Enrico Liscio**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2024

Name of the student: Timur Oberhuber
Final project course: CSE3000 Research Project
Thesis committee: Luciano Cavalcante Siebert, Amir Homayounirad, Enrico Liscio, Jie Yang

Abstract

This study evaluates the performance of different sentiment analysis methods in the context of public deliberation, focusing on hard-, soft-, and subjective-label scenarios to answer the research question: "can a Large Language Model detect subjective sentiment of statements within the context of public deliberation?". If the answer to this question is affirmative, that is a strong indicator that, with the help of longitudinal studies, sentiment analysis with large language models (LLMs) may be implemented to scale public deliberations by providing support for moderators in such discussions. To answer this question, four prompting methods were tested: zero-shot, few-shot, chain-of-thought (CoT) zero-shot, and CoT few-shot using a Frisian dataset of 50 statements annotated by 5 annotators. The findings indicate that the CoT few-shot method significantly outperforms other methods in all scenarios, that soft-labels outperform their hard equivalent, that the underlying data must be balanced for high performing models, and that capturing the perspective of a specific annotator requires further research. Our study suggests that LLMs may perform best under the supervision, or with the collaboration of a human, due to the multi-faced nature of sentiment.

1 Introduction

The increasing urgency of issues like climate change, energy transition, and technological regulation highlights the importance of public deliberation. This drives a growing demand for citizen involvement in decision-making processes [1]. Therefore it is imperative to scale public deliberation, however, an effective deliberation requires at least one moderator per twenty participants [2], as each participant has their own positive or negative attitude towards a topic. This attitude is determined by their "personal feelings, views and beliefs" [3], referred to as "subjective sentiment" in this paper. As a moderator needs to have a good overview of each participant's subjective sentiment, the number of moderators per participant is a large impediment to scaling public deliberation.

To address this limitation, sentiment analysis has been proposed as a tool to support moderators in large-scale public deliberations. It may allow for more participants per moderator by providing moderators with a more detailed overview of each participant [4]. Although there exists a multiplicity of tools for sentiment analysis, the recent emergence of large language models (LLMs) enables novel approaches to sentiment analysis in public deliberations, as will be addressed in Section 3. Furthermore, as will also be addressed in Section 3, existing research and methods focusing on the use of LLMs on sentiment analysis mainly look at the use of zero and few-shot methods, whereas chain-of-thought (CoT) reasoning, a method shown to perform well with LLMs, is largely missing.

As such the main research question this paper will be answering is as follows: **Can a Large Language Model detect subjective sentiment of statements within the context of public deliberation?**

However, in order to simplify the process of answering this main research question, several smaller sub-questions were initially focused on.

Firstly, as annotating the dataset prior to beginning the experiment was necessary, we needed to determine if annotating subjective sentiment and combining multiple perspectives could be effectively achieved in the context of public deliberation. Then, as part of the experiment, each of the following prompting methods was implemented and tested with an LLM model against the test-set in three different scenarios (soft-, hard-, and subjective-label): zero-shot, few-shot, zero-shot CoT, and few-shot CoT. Thus answering the question: **to what extent can zero-shot, few-shot, CoT zero-shot, or CoT few-shot LLMs be implemented to detect subjective sentiment in the context of public deliberations?**

Finally, when evaluating the four methods, accuracy and F1-score were considered in order to answer the question: **how can LLMs be be evaluated over subjective sentiment detection in the context of public deliberation?** This resulted in the conclusion that the CoT few-shot method significantly outperforms other methods in all scenarios and, in turn, the soft-label scenario outperforms all others. This indicated that soft-labels are generally more performant than hard ones due to the multi-faceted nature of sentiment. Finally, it was determined that LLMs would likely perform best in collaboration with, or under the supervision of a human, due to the aforementioned multi-faceted nature of sentiment.

In order to address the findings in this paper, firstly the Background and Related Work are discussed, followed by an explanation of the Methodology. Thirdly, the Results are presented and discussed in the Discussion section. Finally, the paper concludes with an explanation of the Conclusions, Future Work, and Responsible Research.

2 Background

Public deliberation has existed as a concept since the beginning of democracy, however, no specific definition has been centrally agreed upon. As such, in this paper the definition proposed by Blacksher et al. in 2012 will be used, which defines public deliberation as a form of informed, value-based discussion that prioritizes the inclusion of ordinary people, particularly marginalized groups, in finding transformative solutions to challenging social problems [5].

As public deliberation is becoming more important in society, the demand for citizen participation has increased

significantly [1]. As such it is vital to scale public deliberation, however, as mentioned previously, effective deliberation requires at least one moderator per twenty participants [2], which is a large impediment to scaling public deliberation.

The field of sentiment analysis, also known as opinion mining, involves extracting the opinion polarity from a piece of text. In essence, this means classifying whether a piece of text is expressing a positive, negative, or neutral opinion [3]. It has been proposed as a solution to the aforementioned limitation by integrating it into public deliberations as a support for moderators engaged in massive public deliberations [4].

2.1 Labelling Subjective Concepts

A common issue experienced by all sentiment analysis methods, which are discussed in more detail in Section 3, is the collection of labelled data. Due to the subjective nature of sentiment, in a multiplicity of cases it is often necessary to combine the results from multiple human labellers. The question then becomes how to combine these labels.

The simplest option is called majority aggregation and involves assigning one hard-label based on the majority of annotators [6]. However, this method has been shown to lose valuable data, as labels which may not have had the majority, but have had a large quantity of annotations may be classified equally as bad as labels which were chosen by none of the annotators.

Another successful option in this case is a method called soft-labels. This method involves combining the labels of multiple human labellers by averaging them and providing a level of confidence instead of one hard label [7]. For instance, if there are three classes - A, B, and C - and two labellers provided the following classifications: 1, 0, 1, and 0, 0, 1; then the soft label for that row would be: 0.5, 0, 1. Not only does soft-labelling provide an efficient manner for combining subjective labels, but Vyas et al. has shown that it works well in combination with meta-learning, which large language models (LLMs) implement [7].

A third successful option is a method which we will call subjective-label, which involves attempting to capture the perspective of each annotator while predicting labels. According to Zhang et al. [8], the performance of LLMs can often be hampered by biases, however, these can be leveraged in order to enhance model performance, which is what the subjective-label method aims to do.

These three labelling scenarios are significant in the case of this paper, as it focuses on using LLMs to determine sentiment. Namely, this paper aims to analyse the difference between zero-shot, few-shot, and chain-of-thought (CoT) LLM performance on sentiment analysis in the three aforementioned labelling scenarios.

2.2 LLMs

In order to understand the application of LLMs on sentiment analysis, it is first important to grasp the fundamentals of these methods. According to Chen et al., the use of a zero-shot method on an LLM to conduct sentiment analysis involves predicting sentiment labels without providing any training samples in addition to the training data of the "out-of-the-box" model. Generally, LLMs are particularly good when used with the zero-shot approach compared to other machine learning methods, due to the use of "auxiliary information that describes inter-class relationships" that is provided in the "out-of-the-box" model. Similarly, also as explained by Chen et al., the few-shot approach involves building on an "out-of-the-box" model by providing the LLM with a relatively small number of examples or training data prior to asking it to complete a task [9].

CoT reasoning prompting, which can be used in combination with both the zero and few-shot methods described above, involves asking the LLM to provide its reasoning when coming up with its response. In essence, implementing this requires the addition of "let's think step-by-step" to the end of the prompt [10]. As previous research has shown that CoT reasoning can provide a large increase in the performance of LLMs, this method is especially interesting to look into with a focus on the context of sentiment analysis.

3 Related Work

Prior to LLMs, there were three main methods of sentiment classification: machine learning, lexicon-based, and a hybrid of the two [11]. As mentioned previously, the recent emergence of large language models (LLMs) enables novel approaches to sentiment analysis and although LLMs technically fall under the classification of machine learning, quite often they are the result of a combination of machine learning or hybrid methods. For instance, LLMs are combined with traditional sentiment analysis tools in order to extract relevant keywords which the traditional tool then utilizes to determine sentiment. As found by Deng et al., an LLM can extract more stable, and accurate labels than traditional methods, although the final performance is on-par with existing methods [12]. Another study, by Lofty et al., found that using LLM translation as a pre-processing step with traditional methods determining the sentiment of the translated text was effective at determining sentiment in Arabic texts [13].

Another approach is to directly utilize LLMs for sentiment analysis, as in this paper, where a common theme is the comparison of zero-shot and few-shot prompting methods to fine-tuned models. For instance Hasan et al., compared fine-tuned language-specific models to general models using zero- and few-shot approaches, and found that fine-tuned models outperform the general ones at sentiment analysis in uncommonly spoken languages [14]. Conversely, Juros et al,. found that LLMs outperform fine-tuned traditional methods when applied to news headlines [15], while Kuila and Sarkar [16], found that LLMs exhibit potential for sentiment analysis in the domain of political news.

One major gap in the aforementioned research, is the

lack of research into chain-of-thought (CoT) prompting. Although zero-shot and few-shot approaches are thoroughly research, CoT reasoning, which has been shown to perform remarkably well [10] with LLMs, has not. In addition, although sentiment analysis is often mentioned in "future research" or "future work" sections in papers on public deliberation, there are none which explore this thoroughly. The combination of these two knowledge gaps leads to the research conducted in this paper.

4 Methodology

In order to answer the research question, it was decided to conduct an experiment. This section describes the experimental setup and methodology, along with the reasoning behind it, however, a general overview can be seen in Figure 1.

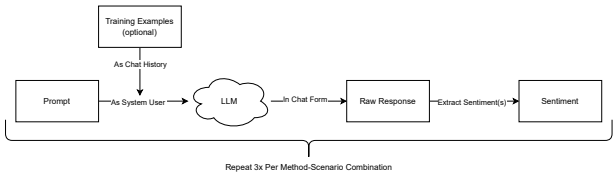


Figure 1: An Overview of the Methodology

4.1 Controlled Variables

To ensure that the experimental results can be compared, some controlled variables needed to be decided on. Namely:

- **LLM Model:** Meta's `Llama 3` model via Ollama [17]. This model was decided on as it was trained on an extensive dataset and fine-tuned for conversational-scenarios [18], but more importantly, it is also audited for safety and bias in the training data. Namely, Meta "performed extensive red teaming exercises, performed adversarial evaluations and implemented safety mitigations techniques to lower residual risks" [18].
- **LLM Temperature:** The temperature setting on an LLM determines the level of randomness and "creativity". When the temperature is set to zero, the responses of the model are reproducible, but generally produce less valuable results [19]. As such we decided to keep the default value of 0.7 for the temperature of our LLM and to run each experiment thrice in order to improve reproducibility, as will be discussed in Section 9.
- **Dataset:** Consisting of the anonymized textual opinions of 1376 residents of Sudwest-Frysland on the future energy policy of their municipality, this dataset comes from a 2020 study by Sprit SL. and Mouter N. from TU Delft [20]. It is important to note that this dataset was translated from Frisian to English, and as such contains some minor grammatical errors and Frisian phrases.

In addition, to ensure a fair experimental setup, the same Python version (3.12.3), Ollama version (0.1.8), and hardware (Windows Machine with an i7-10750H CPU and 16GB of RAM) were utilized across all simulations.

4.2 Labelling the Dataset

Prior to being able to begin any experiments the dataset needed to be labelled.

Firstly, we decided on the possible sentiment categories that a data point could be labelled as. Based on research by Liu, which indicates that a fine-grained approach provides more detailed insights, we decided on a five-class categorization of: Strongly Negative, Slightly Negative, Neutral, Slightly Positive, and Strongly Positive [3].

Due to the availability of peers, it was decided to each label 50 rows of the data independently. This resulted in 50 rows of annotated data, each with the annotations from five different raters. After individually annotating the dataset, the labels were aggregated using three approaches as described below in more detail: hard-label, soft-label, and subjective-label.

Soft-Labels

For each data point, the proportion of annotators who assigned it a particular label was calculated. This proportion was then recorded as the final label value for that data point [7]. For instance, if there are five annotators who annotated a data point with the following labels: strongly negative, slightly negative, neutral, slightly positive, and strongly positive; then the final soft label for that data point would look as described in Table 1.

Table 1: An example of the result of soft labelling data.

Strongly Negative	Slightly Negative	Neutral	Slightly Positive	Strongly Positive
0.6	0.2	0.2	0	0

Hard-Labels

For each data point, the proportion of annotators who assigned it a particular label was calculated. The final "hard" or "correct" label was then chosen based on which label had the highest proportion [6]. In the case of an equal highest proportion, the first seen label was selected.

Subjective-Labels

In this scenario, all the labels were left unaggregated, as the goal of this scenario is to capture the unique labelling perspective of each annotator.

4.3 Experiment

The experiment involved systematically evaluating four distinct prompt engineering techniques, each in the hard-label, soft-label, and subjective-label scenarios. In order to do this, each of the following prompt engineering techniques was applied once on hard-label data, once on soft-label data, and once on subjective-labelled data. Each of the prompt engineering techniques and their respective prompts and minor differences in methodology are described below.

However, prior to evaluating the four methods in each of the scenarios, we split the labelled dataset into training

(70%) and testing (30%) sets. This standard split ensured a robust evaluation of model performance on unseen data with a smaller dataset.

Zero-Shot

Description: This method was chosen as a "base-line" method, as in this approach, the model was asked to directly predict sentiment for each of the text inputs in the test set without any prior task-specific examples. The prompts described below were provided to the LLM via the "system" user. The model was then asked to classify the text from the test set.

Hard-Label Prompt: The prompt for this scenario instructed the LLM that it had the role of a "sentiment analysis model", then provided the available sentiment categories. Finally, the model was instructed to return a number between zero and four to indicate the sentiment of the provided input. The full text of the prompt can be found in Appendix A.1.

Soft-Label Prompt: The prompt for this scenario similarly instructed the LLM that it had the role of a "sentiment analysis model", then provided the available sentiment categories. The prompt then indicated that a probability distribution was to be provided as an answer with some rules describing the format and properties of the distribution (e.g. sum of values must equal to one). The full text of the prompt can be found in Appendix A.1.

Subjective-Label Prompt: Due to the required examples to capture an annotator's unique perspective, the zero-shot method was not applicable in this scenario.

Few-Shot

Description: With the few-shot approach, the model was instructed to predict sentiment for each of the text inputs via the "system" user. It was then provided with the training set as examples of text-sentiment pairs prior to making predictions, in the format of user-assistant conversation history. The model was then asked to classify the text inputs from the training set. As such, this method was used to compare the more commonly researched prompting methodology to the less researched chain-of-thought (CoT) prompting, as described in Section 3.

Hard-Label Prompt: The same prompt was used as described above for the hard-label scenario of the zero-shot approach, the full text for which can be found in Appendix A.1.

Soft-Label Prompt: The same prompt was used as described above for the soft-label scenario of the zero-shot approach, the full text for which can be found in Appendix A.1.

Subjective-Label Prompt: The LLM was explained its role as a "sentiment analysis model", provided the five possible sentiment categories, then instructed to consider each annotator's sentiment annotation history, and to predict each annotator's sentiment annotation for the input. The full version of the prompt can be found in Appendix A.3.

Zero-Shot Chain-of-Thought

Description: The model was guided, via the "system" user, to predict sentiment, then reason step-by-step about its prediction, providing justifications based on identified phrases. With the zero-shot variant of CoT reasoning, the model was not provided any input-sentiment-reasoning tuples prior to being asked to predict the sentiment of the test set. This method was chosen to provide a "baseline" measurement of CoT prompting without any additional training data.

Hard-Label Prompt: The prompt for this scenario instructed the LLM that it had the role of a "sentiment analysis model". It was then instructed to provide two outputs: 1) a value from zero to four indicating the sentiment category, and 2) an explanation of the reasoning behind the chosen sentiment. The full text of this prompt can be found in Appendix A.1.

Soft-Label Prompt: The prompt for this scenario instructed the LLM that it had the role of a "sentiment analysis model". It was then instructed to provide two outputs: 1) a probability distribution, adding up to a total of one, indicating the sentiment category, and 2) an explanation of the reasoning behind the chosen sentiment distribution. The full text of this prompt can be found in Appendix A.2.

Subjective-Label Prompt: Due to the required examples to capture an annotator's unique perspective, the zero-shot CoT method was not applicable in this scenario.

Few-Shot Chain-of-Thought

Description: With this approach the model was guided to predict, and then reason step-by-step about its prediction via the "system" user. The model was then provided text-sentiment-reasoning tuples from the training set via the chat history, prior to being asked to classify the test set. In order to provide the example tuples, each data point / text input in the training set was converted into the following format prior to being added to the chat history:

`Input`: [Textual Data-Point].
`Output`: [Example Output (Number / Python Array)].
`Reasoning`: [Example Reasoning]

Hard-Label Prompt: The same prompt was used as described above for the hard-label scenario zero-shot CoT approach. The full text of the prompt can be found in Appendix A.1.

Soft-Label Prompt: The same prompt was used as described above for the soft-label scenario zero-shot CoT approach. The full text of the prompt can be found in Appendix A.2.

Subjective-Label Prompt: For this scenario the LLM was instructed that it was a "sentiment analysis model" and asked to provide two outputs: 1) Based on the history of each annotator the predicted annotation for that annotator,

and 2) an explanation of the predictions. The full text for this prompt can be found in Appendix A.3.

4.4 Evaluation

Once all the results were collected, they were evaluated on the what are considered to be the two main metrics when evaluating machine learning methods [21]:

1. **Accuracy:** The percentage of correctly classified sentiments. Calculated using the following formula:

   $$\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i)$$

   Where N is the number of data points, $\hat{y}_i$ is the predicted label value for the i-th item, and y_i is the reference/true label value for the i-th item.

2. **F1-Score:** The statistic which indicates the combined performance of precision and recall, calculated using the harmonic mean between precision and recall. Determined using the following formula:

   $$\frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

   With:

   $$\text{Precision} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c}$$

   $$\text{Recall} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}$$

   where C is the number of sentiment categories (5 in our case).

The choice of accuracy and F1 score as evaluation metrics is motivated by their complementary strengths in assessing different aspects of model performance. **Accuracy** is a straightforward and intuitive metric that measures the overall correctness of the model. It is particularly useful when the classes are balanced and gives a quick overview of the model's performance. However, in scenarios where the data might be imbalanced, relying solely on accuracy can be misleading. In such cases, a model could achieve high accuracy by simply predicting the majority class, neglecting the minority class altogether. This is where the **F1-Score** becomes crucial. It provides a single metric that balances both false positives and false negatives. In addition, instead of reporting both precision and recall, we can report one value. By using both accuracy and F1-Score, we can comprehensively evaluate the model's performance, ensuring that it is not only correct overall but also effective in identifying all relevant instances across different classes.

However, in the soft-label scenario, each of the above mentioned metrics needs to be adjusted to become fuzzy. According to Harju et al. [22], it is possible to use fuzzy accuracy and F1-Scores equally as well as their hard equivalents for evaluation in the case of soft-label meta training. They propose the following formulas in order to calculate these values:

1. **Fuzzy Accuracy:** Uses cosine similarity to determine the distance between the two probability distributions and can be calculated using the following formula:

   $$\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{a}_i \cdot \mathbf{b}_i}{\|\mathbf{a}_i\| \|\mathbf{b}_i\|}$$

   Where N is the number of data points, a_i is the predicted label value for the i-th item, and b_i is the reference label value for the i-th item.

2. **Fuzzy F1-Score:** The same as F1-Score, except Fuzzy Precision and Recall are used:

   $$\frac{2 \cdot \text{Fuzzy Precision} \cdot \text{Fuzzy Recall}}{\text{Fuzzy Precision} + \text{Fuzzy Recall}}$$

   With:

   $$\text{Fuzzy Precision} = \frac{\sum_{i=1}^{N} \min(y_i, \hat{y}_i)}{\sum_{i=1}^{N} \hat{y}_i}$$

   $$\text{Fuzzy Recall} = \frac{\sum_{i=1}^{N} \min(y_i, \hat{y}_i)}{\sum_{i=1}^{N} y_i}$$

   where N is the number of data points, $\hat{y}_i$ is the predicted label value for the i-th item, and y_i is the reference label value for the i-th item.

5 Results

This section presents the results of our experiment, focusing on the performance of different methods under hard-label, soft-label, and subjective-label scenarios. We start with an evaluation of annotator reliability and possible bias, before comparing the performance metrics of the methods throughout the scenarios.

5.1 Annotator Evaluation

In order to ensure that the results are reliable and valid, it is first important to determine if the annotations provided for the data are reliable. As such the following factors were considered:

Bias

Firstly, in order to ensure that none of the annotators were biased towards rating more positively or negatively than any of the others, we calculated the average annotation scores (between 0 and 4) for each annotator.

Table 2: The Average Annotations of Raters/Annotators

Annotator	1	2	3	4	5
Avg. Annotation	2.26	2.08	1.9	2.0	2.02

As can be seen in Table 2, all the annotations are within 0.26 of 2, a neutral rating. This is already an indicator that there is no major bias, however, to further ensure this, we ran a one-way ANOVA test, which resulted in a *p-value* of **0.639**. As the *p-value* is significantly higher than 0.05, we can conclude that there is no significant difference in average ratings among annotators, indicating no significant bias among annotators.

Inter-Annotator Agreement (IAA)

Secondly, to ensure that the consistency of annotations between annotators is high; which in turn ensures reliable data to train machine learning models on; we ran a split-half

reliability test using Fleiss' Kappa with 10,000 splits.

Our five annotators resulted in a mean Fleiss' Kappa score of 0.17, indicating low agreement, but no systemic disagreement. However, as the test resulted in a *p-value* of 1.0, this lack of agreement is likely attributable to random variability rather than a systematic difference.

5.2 Comparing the Methods

We now present the performance of different methods under hard-label, soft-label, and subjective-label scenarios. Tables 3a, 3b, 3c summarize the performance metrics (Accuracy and F1-Score) for the different methods under hard-label, soft-label, and subjective-label scenarios respectively.

Table 3: Comparison of Methods by Scenario

(a) Hard-Label Scenario

Method	Accuracy	F1-Score
Zero-Shot	0.143	0.162
Few-Shot	0.286	0.324
CoT Zero-Shot	0.357	0.352
CoT Few-Shot	0.500	0.386

(b) Soft-Label Scenario

Method	Accuracy	F1-Score
Zero-Shot	0.514	0.430
Few-Shot	0.646	0.529
CoT Zero-Shot	0.596	0.476
CoT Few-Shot	0.702	0.593

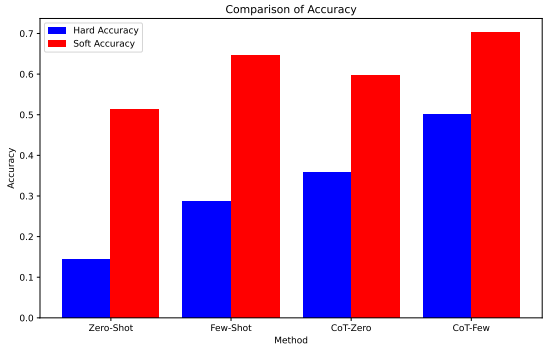
(c) Subjective-Label Scenario by Annotator

Annotator	Few-Shot		CoT Few-Shot	
	Accuracy	F1	Accuracy	F1
1	0.214	0.139	0.357	0.325
2	0.143	0.077	0.500	0.330
3	0.357	0.271	0.357	0.292
4	0.214	0.145	0.357	0.296
5	0.143	0.103	0.429	0.426

Looking at these results, several significant observations can be made. Namely:

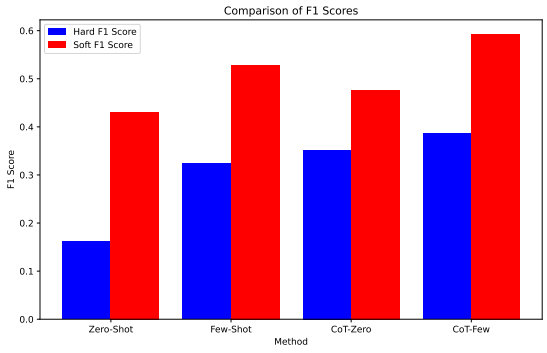
- **Hard-Label Data:** All methods perform relatively poorly, with an accuracy and F1-Score below or equal to 0.5. Notably, F1-Scores are always less than accuracy with the exception of the few-shot method in this scenario. Furthermore, our findings demonstrate a clear trend of performance improvement (around 10% with each consecutive method) with increasing training data and inclusion of reasoning capabilities.
- **Soft-Label Data:** The CoT few-shot method outperforms the other methods in all metrics, although the few-shot method is not too far behind, indicating that although increased training data provides some performance improvement, reasoning capabilities play a larger role. Additionally, we note that the F1-Score is consistently about 0.1 lower than the accuracy.
- **Subjective-Label Data:** In general, the LLM performs relatively poorly at capturing each annotator's perspective, with an accuracy and F1-Score below 0.5 for all annotators and methods. We note that the CoT few-shot method outperforms the few-shot method in accuracy and F1-Score for all annotators except 3, where the accuracy is equal, indicating that reasoning capabilities play a significant role in the capturing of unique perspectives. Notably, for annotator 3, both methods perform identically to three decimal places in terms of accuracy. Finally, the F1-Scores are consistently around 0.1 lower than the accuracy for the few-shot method, but relatively equal or at most 0.5 lower for the CoT few-shot method.

Figure 2: Accuracy in Soft- and Hard-Label Scenarios



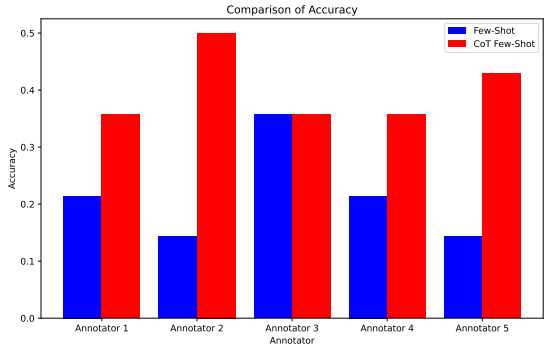
Comparing metrics between hard- and soft-label scenarios in Figures 2 and 3, reveals that all methods perform better with soft-label data. Furthermore, we note that the difference in performance between the hard- and soft-label scenarios becomes smaller the more training data and reasoning capabilities we have. We also notice that although for the hard-label scenario the accuracy and F1-Score increase consistently with each method that adds training data or reasoning capability, in the soft-label scenario the increase is larger when training data is added to non-reasoning methods, than to reasoning methods. In fact, when reasoning is added and training data is removed, the accuracy and F1-Score drop slightly as in the case of few-shot to CoT zero-shot.

Figure 3: F1-Scores in Soft- and Hard-Label Scenarios



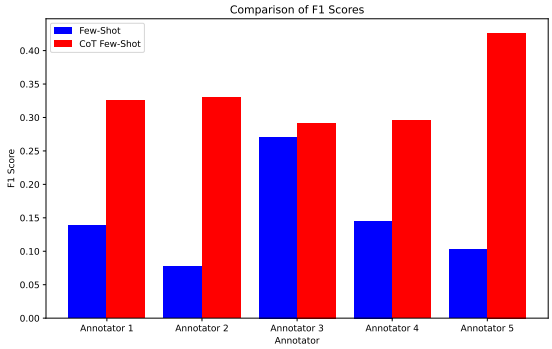
Furthermore, when looking at Figures 4 and 5, we notice that for both few-shot and CoT few-shot methods we get an accuracy and F1-Score similar to the respective scores that those methods achieved in the hard-label scenario. Additionally, for annotator three both methods perform equally, with an accuracy identical to three decimal places.

Figure 4: Acc. Over Raters' Perspectives in Subj. Scenarios



Finally, these results indicate that the answer to our research question, can an LLM detect subjective sentiment of statements within the context of public deliberation, is **yes**, with the general best choice being CoT few-shot.

Figure 5: F1 Over Raters' Perspectives in Subj. Scenarios



6 Discussion

Overall the experiment revealed several important insights into the performance of the four methods across hard-, soft-, and subjective-label scenarios. In this section, we aim to delve into these insights, and to discuss the limitations of this study.

Firstly, we noticed a clear trend of improvement with each method with the addition of training data or reasoning capabilities across all scenarios. However, while we noticed that the addition of either training data or reasoning resulted in a similar improvement in the hard-label scenario, in the soft-label scenario we found that the addition of reasoning capabilities provided a slightly smaller increase than the addition of training data. We suppose, based on this observation, that training or historical data is more valuable than being able to explain the pattern in the data, which makes sense as training data is correct, whereas explanations may still be as flawed being predictions.

Secondly, we note that chain-of-thought (CoT) few-shot outperforms all other methods in all scenarios, except for annotator three in the subjective-label scenario. This is likely due to the fact that this method combines reasoning capabilities with training data, which allows it to reap the benefits of both. This indicates that while each provides a smaller increase in both metrics independently, the combination of both is the most effective method, and that likely the easiest manner of improving accuracy and F1-Score is by providing additional training data with well-reasoned explanations.

Table 4: Balance of Data in Hard- and Subjective-Label Scenarios

Class	Strongly Neg.	Slightly Neg.	Neutral	Slightly Pos.	Strongly Pos.
Count	7	7	11	22	3

Furthermore, in all scenarios we observed that the accuracy is consistently around 0.1 higher than the F1-Score. The reasons for this are likely as follows:

1. **Data Imbalance:** There is a significant imbalance in the underlying distribution of the data as shown in Figure 4 (although these numbers refer to the hard-label and subjective-label scenarios, the soft-label scenario is a probability distribution of this data, and thus has the same underlying imbalance).
2. **Performance on Minority Classes:** For all the methods, we notice that the proportion of mistakes made on the classes which are under-represented in the data; namely classes 0, 1, and 4; is higher than that on majority classes 2 and 3.

In order to resolve this issue it is important to balance the underlying data as will be described in Section 8. Along a similar vein, we also notice that in the subjective-label scenario, although the results are generally similar to those of the few-shot and CoT few-shot methods in the hard-label scenario in terms of accuracy and F1-Score (varying slightly for each annotator); for annotator 3, both methods perform equally to three decimal places in terms of accuracy. This is likely due to a higher internal consistency of annotator 3's annotations compared to the others', making it possible for both methods to perform equally. As described in Section 8, improving inter-annotator consistency may help to improve the consistency between methods for all annotators.

These findings have significant implications for the application of LLMs in sentiment analysis within public deliberation contexts. Namely, the superior performance of the CoT few-shot method in all the scenarios suggests that both reasoning capability and training or historical data are integral parts of an LLM's sentiment analysis skills. In addition, the higher performance of the LLM in the soft-label

scenario indicates that instead of providing moderators with a concrete sentiment, it may be more beneficial to provide a percentage-based overview along with explanations. This approach allows moderators to infer and interpret sentiments more flexibly, aligning with the nuanced nature of public discourse.

Extrapolating from this, we argue that LLMs should not replace human judgment, but rather serve as collaborative tools. Through the insights provided by methods like CoT few-shot, human moderators can be significantly aided by giving them a more detailed understanding of the overall sentiment, which they can then interpret within the broader context.

However, several limitations of our study are important to consider. Firstly, the low inter-annotator agreement, as indicated by the Fleiss' Kappa score, suggests variability in how annotators interpreted the data. This could have impacted the reliability of the training data, potentially affecting the performance of the methods evaluated especially affecting the similarity or dissimilarity of the accuracy and F1-Scores. Secondly, the small number of annotators and the potential for random variability in their ratings could introduce noise into the results. Additionally, in the hard-label scenario, the first seen label was used when percentages were equal. This could have introduced a slight bias to the sentiment categories which appeared earlier in the annotations. Finally, the limited amount of labelled training data could have exacerbated the aforementioned issues, highlighting the importance of a large set of reliably annotated training data.

7 Conclusions

The study aimed to explore whether Large Language Models (LLMs) can effectively detect subjective sentiment in statements made during public deliberation by answering the research question: can a Large Language Model detect subjective sentiment of statements within the context of public deliberation? The primary finding from the experiment and data analysis is that LLMs can indeed detect subjective sentiment within the context of public deliberation. However, the study also emphasizes that LLMs should not be seen as replacements for human judgment. Instead, they should be used as collaborative tools to enhance the process of sentiment analysis.

One of the insights from the study is the success of the soft-label scenario. This suggests that sentiment is not always binary or singular. Instead, it can be complex and layered, making it essential to adopt flexible labeling approaches that capture this complexity. Furthermore, the study also highlights the effectiveness of the chain-of-thought (CoT) few-shot method, which was found to be the best performing method across all scenarios, indicating that the combination of reasoning capabilities and training data is the most effective method for improving sentiment analysis in an LLM.

Moreover, our study highlights the importance of balanced underlying data in improving F1-Score. Balanced data ensures that the model is exposed to a diverse range of sentiments and contexts, preventing it from becoming biased towards certain sentiments.

8 Future Work

Beyond the research described in this paper, there exist a multitude of future avenues to be addressed. Firstly, this paper focused on public sentiment analysis on a Frisian dataset, thus replicating this study with the same methods on a variety of different (possibly larger) datasets and contexts would be beneficial to investigate performance across languages, topics, and cultures, and to provide more generalizable insights.

Secondly, future research could explore methods for improving annotator agreement to enhance the training data provided to models and the ability of large language models (LLMs) to capture the perspective of specific annotators. For instance, this could involve developing more objective annotation guidelines for annotators. This would, in turn, result in a more uniform and easily recognizable annotation style, making it easier for LLMs to capture the perspective of each annotator.

Thirdly, future studies may ensure that the underlying data provided to the LLM is balanced; for instance by over- or under-sampling data, or by generating a synthetic dataset. This would ensure that the LLM would be trained on an unbiased dataset and would improve F1-Scores by ensuring the model is able to predict sentiment equally for each class.

Additionally, further research could be conducted into different techniques involving LLMs and sentiment analysis, especially in the sphere of emulating a specific annotator's perspective. For instance, one interesting method may be fine-turning through reinforcement learning with human feedback (RLHF), as this method may be relatively feasible to implement in a real world scenario, due to a deliberation generally involving one or more moderators who can provide real-time feedback to the LLM. In addition, active learning may be investigated to potentially reduce the annotation effort involved with sentiment analysis. Furthermore, hybrid approaches that combine the strengths of multiple methods to enhance performance across the board may be of interest to research.

Finally, it may be valuable to investigate the effects of providing real-time sentiment analysis to moderators during public deliberations. This could be studied through a longitudinal study wherein an LLM sentiment analysis tool is directly integrated into a deliberation platform. This would allow a more in-depth view of the potential benefits, such as greater participant engagement and inclusivity, as well as into any potential challenges of such an approach.

9 Responsible Research

Firstly, it is important to look at the data that is used throughout the research. Namely, the dataset was sourced from a study conducted by Spruit SL. and Mouter N. from TU Delft in 2020. The dataset consists of the anonymized textual opinions of 1376 residents of Sudwest-Frysland about the future energy policy of their municipality [20]. Although the dataset was previously anonymized by the source of the dataset, it is important to acknowledge the potential for re-identification through the possibility of linking answers between different questions, especially given the specific regional context. Furthermore, it is important to note that although this data was used in this study, it is currently not public and therefore not accessible openly[1].

Although utilizing the `Llama 3` model, with its rigorous evaluation and auditing processes, potential biases in the pre-trained model or the Frisian dataset itself cannot be entirely discounted. These biases could arise from factors such as the under-representation of demographics in the training data (considering the Frisian context) or the specific topic of energy policy. Further research is needed to fully understand and address these potential biases, as addressed in Section 8. In addition, the low number of annotators also represents a risk for bias, as they may not be representative of general public opinion, and as such, may bias the data that the LLM model was trained on.

It is also important to ensure that the research discussed in this paper is reproducible, therefore, not only is the methodology along with the underlying reasoning available in the Methodology section, but the code is open source, and available on GitHub[2], along with a detailed 'ReadMe' document. Furthermore, when generating train/test sets of the data via splits, a random seed of **42** was used to ensure reproducible randomness.

One important limitation in terms of reproducibility was setting the temperature of the LLM. When the temperature is greater than zero, it is not possible to perfectly reproduce the output of the LLM, however, setting the temperature to zero would result in the LLM being unable to determine sentiment due to the lack of insight and creativity [19]. Thus, in order to counter this limitation, for each of the methods and scenarios the experiment was completed thrice, and the results averaged. Thus, the LLM's results are generally, if not perfectly, reproducible.

Finally, when conducting research it is also important to take ethical concerns into consideration. In the case of public sentiment, it is important to consider the possible negative consequences, although there is a multiplicity of positive ones. For instance while this research may result an improved understanding of public opinion and enhanced facilitation for moderators, there are also many possible dangers, such as the potential for the manipulation of results (via a biased model) and an over-reliance on technology to make decisions that affect the public. It is vital to consider the effects of using AI in the decision-making processes that affect the public, as these may result in far-reaching effects that may not be easily reversed. Namely, through the potential misinterpretation of results, existing biases in the public may be reinforced, for instance resulting in the implementation of further biased laws which may disproportionately affect marginalized groups in the context of public deliberation on law.

[1] For access to, or more information regarding the data, please contact the authors of [20].

[2] https://github.com/Timur-O/Research-Project

References

[1] D. Schleifer and A. Diep, "Strengthening democracy: What do americans think," *New York, NY: The*, 2019. [Online]. Available: https://publicagenda.org/wp-content/uploads/Strengthening_Democracy_WhatDoAmericansThinkFINAL.pdf

[2] M. Klein, "Enabling large-scale deliberation using attention-mediation metrics," *Computer Supported Cooperative Work (CSCW)*, vol. 21, pp. 449–473, 2012. [Online]. Available: https://doi.org/10.1007/s10606-012-9156-4

[3] B. Liu, *Sentiment analysis and opinion mining*. Springer Nature, 2022. [Online]. Available: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[4] I. Verdiesen, M. Cligge, J. Timmermans, L. Segers, V. Dignum, and J. van den Hoven, "Mood: Massive open online deliberation platform-a practical application." in *EDIA@ ECAI*, 2016, pp. 4–9. [Online]. Available: https://research.tudelft.nl/files/53016463/b24b06c6588222eae2ca27a54a9cc11ba3a4.pdf

[5] E. Blacksher, A. Diebel, P.-G. Forest, S. D. Goold, and J. Abelson, "What is public deliberation," *Hastings Cent Rep*, vol. 42, no. 2, pp. 14–17, 2012. [Online]. Available: https://doi.org/10.1002/hast.26

[6] V. Prabhakaran, A. M. Davani, and M. Díaz, "On releasing annotator-level labels and information in datasets," 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2110.05699

[7] N. Vyas, S. Saxena, and T. Voice, "Learning soft labels via meta learning," 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2009.09496

[8] Y. Zhang, H. Li, Z. Li, N. Cheng, M. Li, J. Xiao, and J. Wang, "Leveraging biases in large language models: "bias-knn" for effective few-shot learning," 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.09783

[9] J. Chen, Y. Geng, Z. Chen, J. Z. Pan, Y. He, W. Zhang, I. Horrocks, and H. Chen, "Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey," *Proceedings of the IEEE*, vol. 111, no. 6, pp. 653–685, 2023. [Online]. Available: https://doi.org/10.1109/JPROC.2023.3279374

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 24 824–24 837. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[11] D. Alessia, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," *International Journal of Computer Applications*, vol. 125, no. 3, 2015. [Online]. Available: http://dx.doi.org/10.5120/ijca2015905866

[12] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "Llms to the moon? reddit market sentiment analysis with large language models," in *Companion Proceedings of the ACM Web Conference 2023*. New York, NY, USA: Association for Computing Machinery, 2023, p. 1014–1019. [Online]. Available: https://doi.org/10.1145/3543873.3587605

[13] A. Lotfy, K. Saleh, S. Mohamed, J. Lorance, E. Yehia, K. Mohammed, I. AbdAlbaky, M. Fathy, and T. Yasser, "Sentiment analysis for arabic product reviews using llms and knowledge graphs," in *2024 6th International Conference on Computing and Informatics (ICCI)*, 2024, pp. 411–417. [Online]. Available: https://doi.org/10.1109/ICCI61671.2024.10485037

[14] M. A. Hasan, S. Das, A. Anjum, F. Alam, A. Anjum, A. Sarker, and S. R. H. Noori, "Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis," 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2308.10783

[15] J. Juroš, L. Majer, and J. Šnajder, "Llms for targeted sentiment in news headlines: Exploring the descriptive-prescriptive dilemma," 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.00418

[16] A. Kuila and S. Sarkar, "Deciphering political entity sentiment in news with large language models: Zero-shot and few-shot strategies," 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2404.04361

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.09288

[18] A. Zhang, I. E. Ashimine, S. Z. Peter Chng, J. Spisak, J. Shaughnessy, L. de Oliveira, and A. Vaughan, "Responsibility and safety," 04 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md#responsibility--safety

[19] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, "Is temperature the creativity parameter of large language models?" 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2405.00492

[20] S. L. Spruit and N. Mouter, "Energy in súdwest-fryslân," 2020. [Online]. Available: https://www.tudelft.nl/en/tpm/pve/case-studies/energy-in-sudwest-fryslan

[21] K. P. Shung, "Accuracy, precision, recall or f1?" 4 2020. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[22] M. Harju and A. Mesaros, "Evaluating classification systems against soft labels with fuzzy precision and recall," *arXiv preprint arXiv:2309.13938*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.13938

A Detailed LLM Prompts by Method

A.1 Hard-Label Scenario

Without CoT Reasoning

"You are a sentiment analysis model. You will analyze the text given by the user and provide the sentiment of the text from five sentiment categories: Strongly Negative, Slightly Negative, Neutral, Slightly Positive, and Strongly Positive. The output should be a number from zero (0) to four (4), which represents the corresponding sentiment category. Follow these rules: 1) Return only a number from zero four. 2) Do not provide any additional information or text in your response. Example response: 3."

With CoT Reasoning

"You are a sentiment analysis model. You will analyze the text given by the user and provide two outputs: 1) The sentiment of the text from five sentiment categories: Strongly Negative, Slightly Negative, Neutral, Slightly Positive, and Strongly Positive. The output should be a number from zero (0) to four (4), which represents the corresponding sentiment category. Follow this rule: return only a number from zero four. 2) An explanation or reasoning for the results in a separate paragraph. Do not combine the two outputs. The response should be structured as follows: First, a separate explanation paragraph. Then, the number from zero to four representing the sentiment category. Example response: The text contains mixed sentiments with a stronger leaning towards neutrality. There are slight negative and positive sentiments detected with the use of phrases such as "inconvenient", "annoying", and "supporting", but the overall tone is neutral. 2."

A.2 Soft-Label Scenario

Without CoT Reasoning

"You are a sentiment analysis model. You will analyze the text given by the user and provide a probability distribution across five sentiment categories: Strongly Negative, Slightly Negative, Neutral, Slightly Positive, and Strongly Positive. The output should be a Python list of floats (e.g., [0.2, 0.3, 0.4, 0.05, 0.05]), where each element represents the probability of the corresponding sentiment category. The sum of all probabilities must equal 1.0. Follow these rules: 1) Return only the Python list of floats. 2) Ensure the sum of the probabilities equals 1.0. If not, adjust the values proportionally. 3) Do not provide any additional information or text in your response. Example response: [0.2, 0.3, 0.4, 0.05, 0.05]."

With CoT Reasoning

"You are a sentiment analysis model. You will analyze the text given by the user and provide two outputs: 1) A probability distribution across five sentiment categories: Strongly Negative, Slightly Negative, Neutral, Slightly Positive, and Strongly Positive. The output should be a Python list of floats (e.g., [0.2, 0.3, 0.4, 0.05, 0.05]), where each element represents the probability of the corresponding sentiment category. The sum of all probabilities must equal 1.0. Follow these rules: a) Return only the Python list of floats for this part. b) Ensure the sum of the probabilities equals 1.0. If not, adjust the values proportionally. 2) An explanation or reasoning for the results in a separate paragraph. Do not combine the two outputs. The response should be structured as follows: First, a separate explanation paragraph. Then, the Python list of floats. Example response: The text contains mixed sentiments with a stronger leaning towards neutrality. There are slight negative and positive sentiments detected with the use of phrases such as "inconvenient", "annoying" and "supporting", but the overall tone is neutral. [0.2, 0.3, 0.4, 0.05, 0.05]."

A.3 Subjective-Label Scenario

Without CoT Reasoning

"You are a sentiment analysis model. Analyze the user's text and provide the following output: The sentiment of the text categorized into one of five categories: Strongly Negative, Slightly Negative, Neutral, Slightly Positive, and Strongly Positive. Provide an array of five numbers, each representing the predicted sentiment category by five different annotators. Each number should be between zero (0) and four (4), corresponding to the sentiment categories. Consider the provided history to predict each annotator's sentiment annotation for the new text. Format the output as a Python array: [annotator 1, annotator 2, ..., annotator 5]. Example response: [3, 2, 4, 1, 0]."

With CoT Reasoning

"You are a sentiment analysis model. Analyze the user's text and provide two outputs: 1) The sentiment of the text categorized into one of five categories: Strongly Negative, Slightly Negative, Neutral, Slightly Positive, and Strongly Positive. Provide an array of five numbers, each representing the predicted sentiment category by five different annotators. Each number should be between zero (0) and four (4), corresponding to the sentiment categories. Consider the provided history to predict each annotator's sentiment annotation for the new text. Format the output as a Python array: [annotator 1, annotator 2, ..., annotator 5]. 2) An explanation or reasoning for the results in a separate paragraph. Do not combine the two outputs. The response should be structured as follows: First, the explanation paragraph. Then, the array with five numbers representing the predictions of the sentiment category for each annotator. Example response: The text contains mixed sentiments with a stronger leaning towards neutrality. There are slight negative and positive sentiments detected with the use of phrases such as 'inconvenient', 'annoying', and 'supporting', but the overall tone is neutral. Annotator 1 generally leans towards a slightly negative skew, whereas Annotator 2 skews positive. The other annotators all skew neither way, but 3 and 5 always have the same results. [1, 3, 2, 3, 2]."

B Use of Large Language Models

Outside of the use of LLMs for the experimental aspect of this study, the LLMs ChatGPT and Google Gemini were used to aid in the writing and experimental process. They were used for the following, with the indicated prompts:

- **Suggesting Topics to Include in Sections:** *"I am writing a research paper on the use of LLMs to conduct sentiment analysis on public discourse. This research includes an experiment, in addition to some literature review. I am now working on the [section name] section. Please provide a list of topics, along with a short description, in the form of bullet points that I may include in this section."*
- **Suggesting Improvements to Written Sections:** *"I am writing a research paper on the use of LLMs to conduct sentiment analysis on pubic discourse. This research includes an experiment, in addition to some literature review. I am now working on the [section name] section. I have written the following: [text written by me]. Please provide a list of improvements for this section in the format of a list of bullet points. Please provide explanations and concrete examples for improvement."*
- **Fixing Issues with Code:** *"I have the following piece of Python code. It is attempting to [insert purpose of code]. However, I get the following error: [insert error]. Please explain how I can fix this. Provide concrete explanations and code examples. [insert code]."*