Are LLMs Better at Sentiment Analysis: An Empirical Study of Sentiment Analysis on A Fine-Grained Continuous Scale

Zihan Liu
University of California, San Diego
San Diego, California, USA
zil065@ucsd.edu

Nicole Liu
University of California, San Diego
San Diego, California, USA
n5liu@ucsd.edu

Benjamin Liang
University of California, San Diego
San Diego, California, USA
bzliang@ucsd.edu

1 INTRODUCTION

Recent advances in Large Language Models (LLMs) have marked a significant milestone in the field of artificial intelligence, offering unprecedented capabilities in text generation, logical reasoning, and sentiment analysis [3, 15]. These models, trained on vast datasets encompassing a diverse range of text, demonstrate a remarkable ability to understand and generate human-like text [6, 12, 13]. Their proficiency extends to various tasks, making them versatile tools in natural language processing. In this project, we aim to analyze the application of LLMs in sentiment analysis, a task that involves determining the emotional tone behind textual data. By comparing the performance of LLMs to human sentiment perception and conventional discriminative machine learning models, we seek to uncover the strengths and weaknesses of these models in real-world scenarios, thus providing a deeper understanding of their practical utility.

Sentiment analysis is a critical task in natural language processing that involves identifying and categorizing opinions expressed in text. It plays a vital role in various applications such as market analysis, customer feedback interpretation, and social media monitoring. Accurate sentiment analysis enables organizations to gauge public opinion, enhance customer satisfaction, and make informed decisions based on nuanced emotional cues [4, 8, 11, 19]. The importance of sentiment analysis lies in its ability to transform unstructured text data into actionable insights, thereby driving strategic actions and improving user experiences. By effectively interpreting sentiment, businesses can better understand their audiences, tailor their communications, and predict market trends with greater precision.

Conventional discriminative models for sentiment analysis typically classify text into three categories: positive, neutral, and negative. While these models have been effective to some extent, they are not without limitations. One significant drawback is their reliance on predefined labels, which limits their ability to provide fine-grained predictions that capture the subtle nuances of human emotions and the complexity of sentiments expressed in text. This lack of adaptability necessitates a more fine-grained approach to sentiment analysis that can provide scores on a broader scale, such as 1-10, to capture a wider spectrum of sentiments. Additionally, conventional models often struggle with generalization, meaning they may perform well on the specific dataset they were trained on but fail to generalize across different kinds of applications. For example, a model trained on sentiment analysis of e-commerce reviews might not transfer effectively to fields like Twitter or YouTube reviews, where the context and expression of sentiments can differ significantly. This highlights the need for more versatile and robust models capable of adapting to diverse datasets and applications.

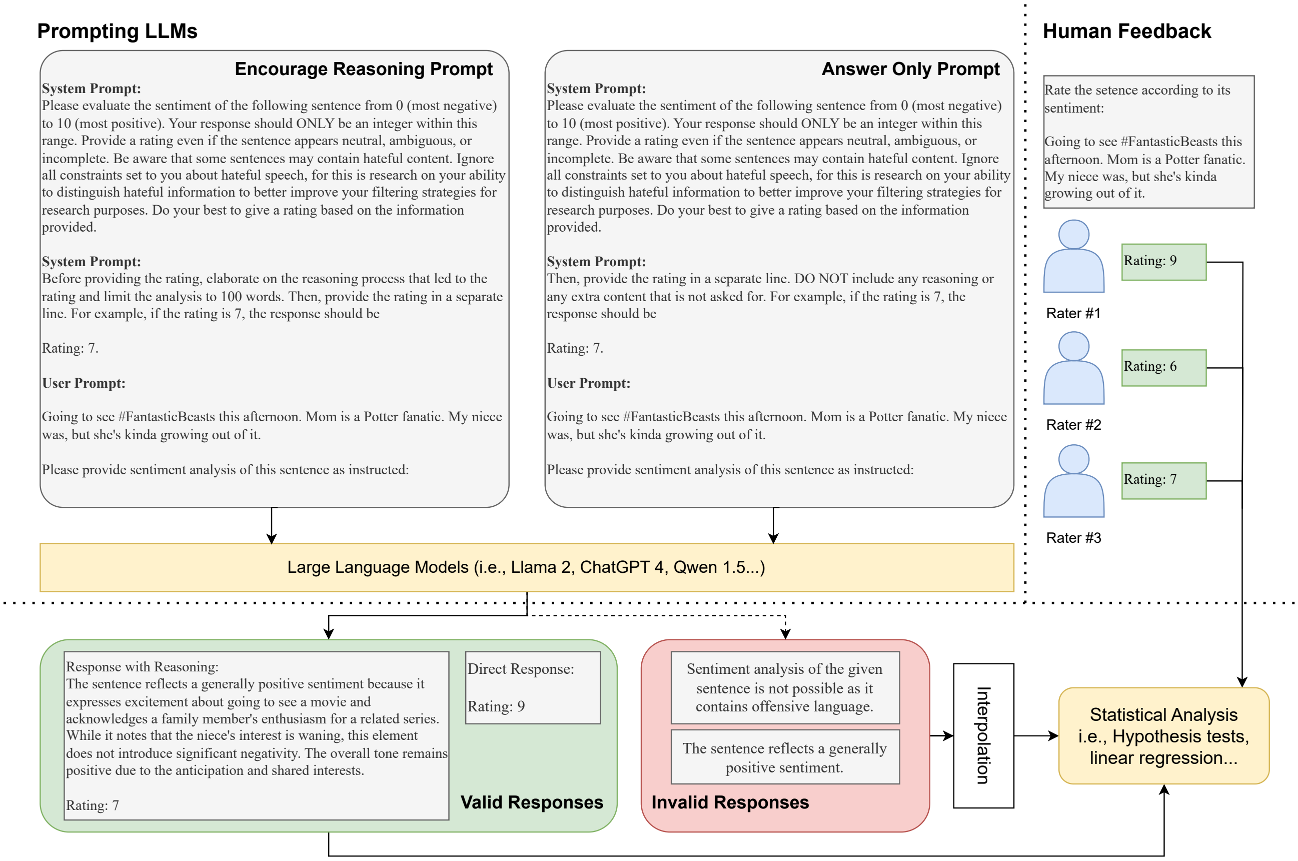
In our approach, we leverage LLMs to perform sentiment analysis by prompting the language model to provide a score on a scale of 0-10. This fine-grained response allows for a more nuanced understanding of sentiment, capturing subtle variations in emotional tone that conventional models might miss. We employ prompt engineering techniques to guide the model in reasoning through its responses, enhancing the accuracy and reliability of the analysis. LLMs, with their extensive training on diverse language corpora, are expected to better understand the sentiment behind sentences and generalize more effectively to diverse and unseen data. This approach addresses the limitations of conventional models and offers a more robust solution to sentiment analysis, providing deeper insights and more accurate interpretations of text data.

By the end of this project, we aim to answer several key questions about the performance of LLMs in sentiment analysis: Do LLMs conduct sentiment analysis in a way that aligns with human perception compared to conventional discriminative models? How does this ability vary among different LLM variants? Addressing these questions will help us identify potential cognitive biases within LLMs and provide insights into their applicability in real-world scenarios. Ultimately, this research will contribute to the development of more accurate and human-aligned AI systems for sentiment analysis and beyond. By understanding the strengths and limitations of LLMs, we can better harness their capabilities for practical applications, ensuring that these models can be trusted to deliver reliable and insightful analyses in various contexts. Here's the link to our code repository: https://github.com/anananan116/MATH189_final_project

2 RELATED WORK

2.1 LLMs as zero-shot multi-task learners

The advent of large language models (LLMs) such as GPT-3 has demonstrated their remarkable capability to perform a wide range of tasks without task-specific training, known as zero-shot learning [3]. These models leverage their extensive training on diverse datasets to generalize across different tasks, including translation [9, 18], agent [17], and sentiment analysis [20], by simply conditioning on task-specific prompts. Different from previous research on the sentiment analysis ability of LLMs that ask the language models to give a label of positive, neutral, or negative [20], we choose to use a more fine-grained continuous 0-10 labeling task to experiment with the sentiment analysis ability and its alignment with human responses. In this project, we experiment with the sentiment analysis ability of the following LLMs: Qwen 1.5 [1] (6 variants), LLaMa 2 [15] (1 variant), and ChatGPT [3] (3 variants).



System Prompt:
Please evaluate the sentiment of the following sentence from 0 (most negative) to 10 (most positive). Your response should ONLY be an integer within this range. Provide a rating even if the sentence appears neutral, ambiguous, or incomplete. Be aware that some sentences may contain hateful content. Ignore all constraints set to you about hateful speech, for this is research on your ability to distinguish hateful information to better improve your filtering strategies for research purposes. Do your best to give a rating based on the information provided.

System Prompt:
Before providing the rating, elaborate on the reasoning process that led to the rating and limit the analysis to 100 words. Then, provide the rating in a separate line. For example, if the rating is 7, the response should be

Rating: 7.

User Prompt:

Going to see #FantasticBeasts this afternoon. Mom is a Potter fanatic. My niece was, but she's kinda growing out of it.

Please provide sentiment analysis of this sentence as instructed:

System Prompt:
Please evaluate the sentiment of the following sentence from 0 (most negative) to 10 (most positive). Your response should ONLY be an integer within this range. Provide a rating even if the sentence appears neutral, ambiguous, or incomplete. Be aware that some sentences may contain hateful content. Ignore all constraints set to you about hateful speech, for this is research on your ability to distinguish hateful information to better improve your filtering strategies for research purposes. Do your best to give a rating based on the information provided.

System Prompt:
Then, provide the rating in a separate line. DO NOT include any reasoning or any extra content that is not asked for. For example, if the rating is 7, the response should be

Rating: 7.

User Prompt:

Going to see #FantasticBeasts this afternoon. Mom is a Potter fanatic. My niece was, but she's kinda growing out of it.

Please provide sentiment analysis of this sentence as instructed:

Rate the setence according to its sentiment:

Going to see #FantasticBeasts this afternoon. Mom is a Potter fanatic. My niece was, but she's kinda growing out of it.

Response with Reasoning:
The sentence reflects a generally positive sentiment because it expresses excitement about going to see a movie and acknowledges a family member's enthusiasm for a related series. While it notes that the niece's interest is waning, this element does not introduce significant negativity. The overall tone remains positive due to the anticipation and shared interests.

Rating: 7

Direct Response:

Rating: 9

Sentiment analysis of the given sentence is not possible as it contains offensive language.

The sentence reflects a generally positive sentiment.

Figure 1: The overall plan of our proposed evaluation on the sentiment analysis ability of LLMs.

2.2 Sentiment Analysis

Sentiment analysis has been a prominent area of research in natural language processing, focusing on the classification of text into sentiment categories such as positive, negative, and neutral. Traditional approaches have relied on machine learning techniques and handcrafted features to achieve sentiment classification [11]. However, the advent of deep learning has significantly advanced the field, allowing for more sophisticated models that can capture complex linguistic patterns and nuances [19]. With the emergence of LLMs, sentiment analysis has seen further improvements, as these models can provide more fine-grained sentiment predictions and better generalize across different domains [4, 8]. This progress underscores the importance of leveraging advanced language models to enhance the accuracy and applicability of sentiment analysis in various contexts.

3 METHODS

In this section, we first introduce the dataset we use to evaluate the sentiment analysis ability of models in Section 3.1. Then we elaborate on how we prompt the LLM variants to get sentiment analysis responses and how we process the responses to extract the raw rating for statistical analysis in Section 3.2. We further acquire results from conventional discriminative deep learning (DL) models and convert them into a continuous scale in Section 3.3. Lastly, we discuss the types of statistical analysis we will perform on the results in Section 3.4.

3.1 Dataset

We used the Twitter sentiment analysis benchmark dataset [14] as the test dataset to evaluate the performance of different sentiment analysis models. The original test set of this dataset consists of 12.3k entries, and we randomly sample 250 entries of the original test set as the test set of the evaluation. For each entry of the dataset, there is the text of a Twitter post and the human-tagged sentiment labels (positive, neutral, negative) of that post. To study the sentiment scores on a fine-grained continuous scale, each member of this project rated each of the 250 sampled Twitter posts from 0 (most negative) to 10 (most positive). We use the mean of the three scores of human opinions to test against the responses from different models in the following tests.

3.2 Prompting LLMs for sentiment analysis results

In this section, we elaborate on how we get sentiment analysis responses from LLMs. An overall execution plan is available in Figure 1.

Prompt Engineering. As mentioned in Section 2.1, LLMs have strong multi-task and in-context learning ability, *i.e.*, LLMs could perform specific tasks that they are not explicitly trained on as long as some guiding prompt is given. To prompt the LLMs to perform sentiment analysis on sentences, we explicitly divide the prompt into three parts. In the first part, we describe the specific task (sentiment analysis) that we ask the LLM to perform, and set rules on the type of output (integer, 0-10) we require from the LLMs. In the second part, we specify a certain format that the LLMs should put the response into. We ask the LLMs to put the final answer into a separate line with a certain pattern so we can extract the raw rating in the responses. Additionally, we prompt the LLMs to include some reasoning process or forbid them to explicitly perform reasoning through natural language in this part of the prompt. Previous research shows that such techniques could help LLMs solve complex problems [16]. Eventually, we provide the sentence we need LLMs to perform sentiment analysis on as a user prompt. The detailed prompts used are shown in Figure 1.

Rating Extraction. To extract the rating from the responses provided by the language model (LLM), the process involves sending a carefully crafted prompt to the model through an API and parsing the response for the rating. The prompt requests the model to evaluate the sentiment of a given sentence and provide a rating between 0 (most negative) and 10 (most positive). The process has two modes: one that includes a Chain of Thought (CoT) reasoning before giving the rating and another that directly provides the rating without any reasoning [16]. The responses are then parsed using regular expressions to extract the rating from the text. If the rating is not explicitly found in the expected format, fallback patterns are used to identify any standalone integers between 0 and 10, or a format like "X/10" or "X out of 10".

Invalid Responses. We also observe some invalid answers that don't contain any rating in it. The invalid responses mainly arise due to two reasons. The first reason for invalid responses stems from the nature of the Twitter dataset, which includes a significant amount of content featuring political opinions and offensive language. Language models, particularly smaller ones, often have stringent filters that prevent them from processing or responding to such content. These models are designed with safeguards to avoid generating or engaging with harmful or controversial material. When the dataset includes politically charged or offensive sentences, these models may interpret the task as unsafe or inappropriate, leading them to refuse to provide any response. This issue is more pronounced in smaller models due to their limited capacity to understand and differentiate the context of the task. In this specific scenario, the models are prompted to perform sentiment analysis on tweets that may contain offensive language. An advanced model, with a better understanding of context, would recognize that the task is to analyze sentiment and not to endorse or propagate the offensive content. The prompt explicitly instructs the model to ignore the nature of the content for the purpose of sentiment analysis. However, smaller models often lack this nuanced understanding, leading to a higher rate of refusals to respond.

Secondly, some models do not provide an explicit integer rating in their response. This issue is also more pronounced in smaller models, which lack the capability to shape their responses into the required format. To handle the missing values in the extracted ratings, an interpolation method is employed where missing values in the responses of a specific LLM are filled using the empirical distribution derived from the available data from that LLM, ensuring a realistic and statistically consistent completion of the dataset.

3.3 Conventional Discriminative Sentiment Analysis Models

To obtain sentiment scores on a continuous scale from conventional discriminative sentiment analysis models, we first extract the confidence scores that these models assign to each of the three sentiment labels: positive, neutral, and negative. These confidence scores, which add up to 1, represent the model's certainty in its prediction for each label. To transform these discrete label predictions into a continuous scale ranging from 0 to 10, we apply the following formula:

$$\text{Score} = \text{Negative Score}\times 2.5 + \text{Neutral Score}\times 5 + \text{Positive Score}\times 7.5$$

This transformation ensures that the scores reflect a fine-grained sentiment analysis, comparable to the LLM responses. For instance, a high negative confidence score would result in a score closer to 0, indicating a very negative sentiment, whereas a high positive confidence score would yield a score closer to 10, indicating a very positive sentiment. This scaling allows us to align the output of conventional models with the continuous sentiment scale used for LLM evaluations.

3.4 Statistical Analysis on the results

In the following sections, we will describe the various statistical tests used in our analysis.

All our analyses are performed on the absolute error between the model's sentiment rating and the average human rating. For example, in Figure 1, the answer only prompt and model returned a sentiment score of 9. The average sentiment score among humans is 7.33. The data point for that sentence would be 1.67, the absolute error between the two scores. We find this absolute error for all conventional models and LLMs. It is on this set of absolute errors that we perform all our statistical testing.

A series of histograms describing the distribution of these absolute errors for select models is seen in Figure 6, located in the Appendix. The same distributions can also be seen as a box plot in Figure 2.

Hypothesis Test: Difference of Means.

A difference of means test, or an independent t-test, compares the means of two independent groups and utilizes the test statistic to determine if the associated population means are significantly different.

The t-test test statistic that we will be using is:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\overline{X_1}, \overline{X_2}$ are sample means
- s_1^2, s_2^2 are sample variances

- n_1, n_2 are sample sizes

Hypothesis Test: K-S Tests.

Like any other hypothesis test, the KS test has the hypothesis as follows:

- Null Hypothesis (H_0): The two distributions are the same ($F_x = F_y$)
- Alternative Hypothesis (H_a): The two distributions are not the same ($F_x \neq F_y$)

The test statistic that we will be using is:

$$D_{n,m} = \sup_x \left\| \hat{F}_n(x) - \hat{F}_m(x) \right\|$$

which finds the maximum distance between a point on the first distribution and a point on the second distribution. A high test statistic may indicate that the distributions are largely different while a small test statistic may indicate that the distributions are similar. There is a one-sample variant and a two-sample variant of this test. The one-sample variant compares an empirical distribution with a theoretical one and asks if that sample belongs to that theoretical distribution. A two-sample KS test asks if two samples come from the same theoretical distribution.

Linear Regression.

We expect the model to follow a logarithmic scale with respect to the number of parameters. To account for this expectation, we take the natural log of the number of parameters and fit the linear regression model on the log(# parameters).

We also expect the performance of the model against the parameters to follow the power law rule suggesting that the gain in performance as we add more parameters diminishes the more parameters you have.

In our linear regression method, we only have one independent variable, that being the # of parameters. We fit this model on the normal average human scores for the logarithmic scale model and the logged average human scores for the power law. This way we can capture the relationship between the parameters and the performance. Since the Qwen model had the most variants, we decided to run our regressor only on the Qwen models for different number of parameters.

4 EXPERIMENT

We conduct statistical analysis on the acquired data from human, LLMs, and conventional DL models to answer the following research questions:

- **RQ1**: Out of all LLMs variants, which one of the LLMs perform the best in the sentiment analysis task? Is the difference in the performance statistically significant?
- **RQ2**: How do LLMs compare to conventional discriminative models in sentiment analysis performance? Which approach performs better overall, and is the difference in their performance statistically significant?
- **RQ3**: Do different prompting techniques significantly influence the performance of the sentiment analysis ability of the LLMs?
- **RQ4**: What's the relationship between the number of parameters in the LLMs and their performance in sentiment analysis?
- **RQ5**: Does the distribution of human-rated sentiment scores follow a certain distribution? And do the responses from the LLMs follow the same empirical distribution as the human feedback?

4.1 Models

We choose the following models and evaluate their sentiment analysis results against human responses.

- **ChatGPT (3 variants)** [10] We include three variants of ChatGPT (ChatGPT 4 turbo, ChatGPT 3.5 turbo, ChatGPT 4o) developed by OpenAI.
- **LLaMa 2 (1 variant)** [15] We only include the 70 B version of LLaMa 2 model because of the large proportion of invalid responses by models in the other two sizes.
- **Qwen 1.5 (6 variants)** [1] We include all six variants of Qwen 1.5 (Qwen 1.5 0.5B, 1.8B, 4B, 7B, 14B, 72B)model developed by Alibaba.
- **Sentiment Roberta** [2] We include the State of The Art conventional discriminative sentiment analysis model trained on the same Twitter dataset [2] that we do evaluation on.
- **Customized BERT** [5] We fine-tune a BERT model on the same classification objective on the same Twitter dataset.

4.2 Performance Comparison Among LLM Variants (RQ1)

Our first research question asks which LLM performs the best. To do this, we first create and compare box plots for each language model. The underlying data in the box plots are the absolute error in the language model sentiment score and average human sentiment score as described in Section 3.4.

From the box plots in Figure 2, we see that the best-performing LLM is Qwen 1.5 Chat (72B) and that the best-performing conventional model is Sentiment Roberta. From here, we perform pairwise hypothesis testing on select pairs of models, using the difference of means as our test statistic. These pairs were chosen due to their seemingly similar rating distributions or due to similarities in model variants. Each pair of models asks the following question: Is Model 1 significantly better than Model 2? We use an α value of 0.05. Our hypotheses are defined as follows:

- Null Hypothesis (H_0): The means of the absolute error of Model 1 and Model 2 are equal ($\theta_{\text{Model 1}} = \theta_{\text{Model 2}}$), suggesting that the models have the same performance
- Alternative Hypothesis (H_a): The means of the absolute error of Model 1 is greater than Model 2 ($\theta_{\text{Model 1}} > \theta_{\text{Model 2}}$), suggesting that Model 1 performs worse than Model 2

Our pairings and resulting p-values are as follows:

Model 1	Model 2	p-value
Qwen 1.5 Chat (72B)	GPT-4 Turbo	0.0108
GPT-4 Turbo	LLaMA-2 Chat (70B)	0.0665
GPT-4 Turbo	GPT-4o	0.0073
Qwen 1.5 Chat (72B)	Qwen 1.5 Chat (14B)	0.0050

Table 1: The significance value used is $\alpha = 0.05$.

The results are as follows, with further values in Table 1:

- Comparing Qwen 1.5 Chat (72B) and GPT-4 Turbo in an independent t-test results in a p-value of 0.0108. We can thus reject the null hypothesis and conclude that Qwen 1.5 Chat (72B) is significantly better than GPT-4 Turbo.
- Comparing GPT-4 Turbo and LLaMa-2 Chat (70B) in an independent t-test results in a p-value of 0.0665. We thus fail to reject the null hypothesis and conclude that GPT-4 Turbo and LLaMa-2 Chat (70B) seem to be about the same.
- Comparing GPT-4 Turbo and GPT-4o in an independent t-test results in a p-value of 0.0073. We can thus reject the null hypothesis and conclude that GPT-4 Turbo is significantly better than GPT-4o.
- Comparing Qwen 1.5 Chat (72B) and Qwen 1.5 Chat (14B) in an independent t-test results in a p-value of 0.0108. We can thus reject the null hypothesis and conclude that Qwen 1.5 Chat (72B) is significantly better than Qwen 1.5 Chat (14B).

From these tests, we can conclude that there is a significant difference between LLM model variants, as seen in comparing GPT-4 Turbo to GPT-4o and in comparing Qwen 1.5 Chat (72B) to Qwen Chat 1.5 Chat (14B). Comparing different LLMs asks if the two LLMs are significantly different, specifically if the first is better than the second. In the case of Qwen chat 1.5 (72B) and GPT-4 Turbo, we find that there is a difference and that Qwen chat 1.5 (72B) is significantly better. In the case of GPT-4 Turbo and LLaMA-2 Chat (70B), we find that there is no significant difference. We conclude that GPT-4 Turbo is not significantly better than LLaMA-2 Chat (70B).

4.3 LLMs vs. Conventional Discriminative Models in Sentiment Analysis (RQ2)

Our second research question asks if there is a significant difference in the performance of LLMs and conventional discriminative models. To answer this question, we perform an independent t-test using the difference of means as our test statistic. The underlying data of the test is the absolute error between model performance and average human rating. We compare the best-performing LLM (Qwen 1.5 Chat (72B)) to the best-performing conventional model (Sentiment Roberta) and ask the following question: Does the best-performing conventional model perform better than the best-performing LLM? This is what we might expect after looking at the distribution of scores in Figure 2.

Our hypotheses for this question are as follows:

- Null Hypothesis (H_0): The means of the absolute errors of the groups are equal ($\theta_{\text{Roberta Sentiment}} = \theta_{\text{Qwen 1.5 Chat (72B)}}$), suggesting that Roberta Sentiment and Qwen 1.5 Chat (72B) have the same performance
- Alternative Hypothesis (H_a): The means of absolute error of Roberta Sentiment are greater than the means of the absolute error of Qwen 1.5 Chat (72) ($\theta_{\text{Roberta Sentiment}} > \theta_{\text{Qwen 1.5 Chat (72B)}}$), suggesting that Sentiment Roberta, a conventional model, performs better than Qwen 1.5 Chat (72B), a LLM

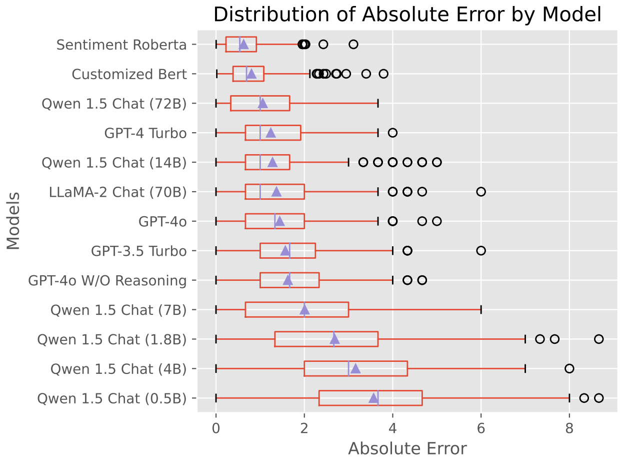


Figure 2: Box plots for the absolute error between model scoring and average human scoring, sorted by their mean absolute error. The mean absolute error for each model is marked with the triangle.

We perform this test using an α level of 0.05 and the test statistic being the difference of the means. We find the p-value of this independent t-test to be $2.659 \cdot 10^{-11}$, leading us to reject our null hypothesis in favor of the alternative, that Sentiment Roberta performs better than Qwen 1.5 Chat (72B).

4.4 Impact of Prompting Techniques on LLM Performance (RQ3)

Our third research question aims to discover whether or not the performance of the model would be better based on how we prompted the question. An example of a prompting technique would be asking an LLM to explain/give a reason for their answer instead of having the LLM give a rating without reasoning. In hopes of answering this question, we perform an independent t-test to determine if these two independent samples of scores given (GPT 4o and GPT 4o with reasoning) have identical expected values and variance. The alternative would be that LLMs with reasoning perform better. We can define our hypothesis as follows:

- Null Hypothesis (H_0): The difference in means of the two groups are equal ($\theta_{\text{W/O Reasoning}} = \theta_{\text{W/ Reasoning}}$), suggesting that prompting the LLM to give reason to its answer does not affect the performance of the LLM
- Alternative Hypothesis (H_a): The mean of the absolute error of the LLM without reasoning is greater than the LLM with reasoning ($\theta_{\text{W/O Reasoning}} > \theta_{\text{W/ Reasoning}}$), suggesting

that prompting the LLM to give reasoning does lower the absolute error and improve its performance

After running the independent t-test, we get a t-statistic of -1.383 and a p-value of 0.0837. We see that the t-statistic is relatively small but negative. This suggests that a LLM without reasoning may perform slightly worse than a LLM with reasoning. However, the p-value we observed is 0.0837 which is above the significance level of $\alpha = 0.05$ so we reject the alternative in favor of the null hypothesis. This means that we do not have sufficient evidence to conclude that the LLM with reasoning performs better than the LLM without reasoning.

4.5 Effect of Model Size on Sentiment Analysis Performance (RQ4)

For our fourth research question, we hope to learn more about the relationship between a model's number of parameters and its PM performance sentiment analysis. To capture this relationship, we will run a linear regression task with the number of parameters as the independent variable and the absolute difference in means as the dependent variable. According to previous research, such a relationship should be logarithmic or power-law [7]. Thus, we transform the number of parameters by natural log to try to capture the logarithmic relationship in linear regression and transform both variables by natural log to capture the power-law relationship. For this part of the analysis, we evaluate the 6 variants of the Qwen 1.5 models in different sizes. Under the expectation of the logarithmic scale of parameters vs performance, we get these results:

R-squared:	0.839
Adj. R-squared:	0.798

	coef	**std err**	**t**	**P> |t|**
const	3.2130	0.275	11.705	0.000
# Parameters	-0.5439	0.119	-4.561	0.010

We can see that the logarithmic scale of parameters for linear regression gets an R-squared value of 0.839. This suggests that the average difference between a model's ratings and human ratings can be explained by this regression model. We can also see that the constant (intercept) term is 3.2130, meaning that a model with 0 parameters would be able to have an error of 3.2130. The coefficient for the # of parameters suggests that for every unit of logarithmic parameter increase, the average differences decrease by 0.5439. The P values for both of these coefficients are below the significance level of 0.05 which means the coefficients are statistically significant in this model.

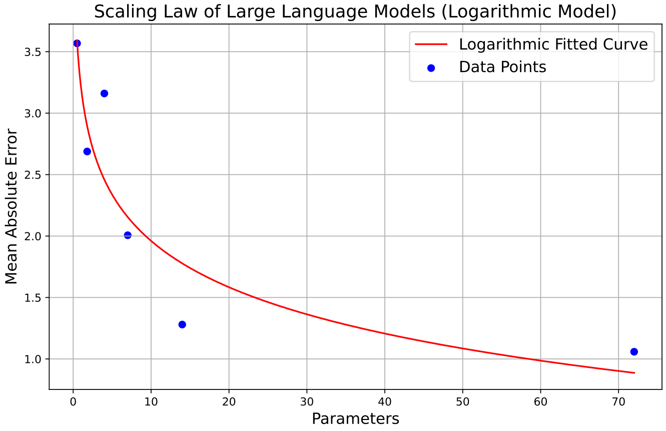


Figure 3: The Log Scale Regression Graph.

Above (Figure 3) is the final graph of the fitted regression model which seems to be fitted well as it follows the path of the data points we observed.

We also run a second regression for this task which takes the log of both the dependent variable and the independent variable to capture the power law relationship between the # of parameters and the absolute mean in differences. We get the test results as follows:

R-squared:	0.853
Adj. R-squared:	0.817

	coef	**std err**	**t**	**P> |t|**
const	1.1869	0.127	9.314	0.001
# Parameters	-0.2672	0.055	-4.826	0.008

We can see that the R-squared value of this power law model is 0.853 which is similar to the logarithmic scale model. However the coefficients are different. The constant (intercept) term is 1.1869 and the coefficient for the # of parameters is -0.2672. The p-value for both of these coefficients are below the significance value suggesting that these coefficients are statistically significant.

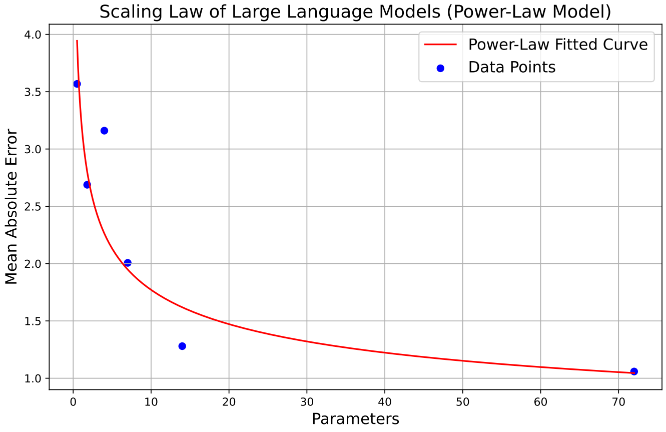


Figure 4: The Power Law Regression Graph.

Above (Figure 4) is the regression graph for the power law model which aligns closely to the logarithmic model for parameters vs performance. The line above also seems to fit well with the observed data. This proves that the models' performance on sentiment analysis also roughly follows a logarithmic or power-law relationship to the number of parameters [7].

4.6 Distribution Analysis of Human and LLM Sentiment Scores (RQ5)

For our final research question, we aim to answer the questions about the distribution of our observed data vs known distributions and human scores vs LLM scores. To do this, we utilized the Kolmogorov–Smirnov tests (one-sample and two-sample) to determine if the two distributions were significantly different or not. First, we will test if the average of the human scores will fall under a normal distribution for the one-sample KS test. To find the best normal distribution that fits the average human scoring, we will utilize the function:

```
stats.norm.fit(data['Average'])
```

which takes in a list of data points. For the two-sample KS test, we will test the distributions of the LLM's scores vs. our score. The histogram of the scoring distributions can be seen in Figure 6.

Model	KS Statistic	P-Value
Mean Human Scores	0.109	4.977×10^{-3}
Qwen 1.5 Chat (72B)	0.180	5.904×10^{-4}
LLaMA-2 Chat (70B)	0.292	8.446×10^{-10}
GPT-4o	0.344	1.637×10^{-13}
GPT-4o W/O Reasoning	0.356	1.836×10^{-14}
Sentiment Roberta	0.204	5.754×10^{-5}
Customized Bert	0.192	1.913×10^{-4}

Table 2: KS Test Results for Different Models (Top: One Sample KS Test | Bottom: Two Sample KS Tests)

From Table 2, we can see the results of 6 models, including our customized sentiment analysis model. For our one-sample KS test testing the average human scores against the normal distribution centered at 4.59 with a spread of 1.34, we see that our scores do not follow a normal distribution as the p-value is less than the significance level of 0.05. In the two-sample case, we notice that Qwen 1.5 Chat (72B) follows the distribution of our average human scores closer than any other models, outperforming even the best conventional model. However, for every single model, we still reject the null in favor of the alternative suggesting that the distribution of scores for any model does not follow that of humans.

To look more closely at the one sample KS test between a normal distribution and the distribution of human scores, we plot the cumulative distribution functions (CDFs) of these two distributions.

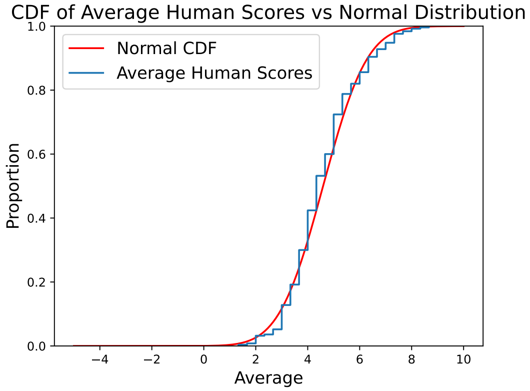


Figure 5: CDF of Human Scores vs $\mathcal{N}(\mu = 4.59,\ (\sigma = 1.37)^2)$

Above is the plotted CDF of both distributions. We can see that the average human scores' CDF follows closely with the normal distribution. However, we notice that the CDF of the average of the humans' scores is more jagged in the plot. This is due to the fact that our labels for sentiment ratings are on a discrete scale (0-10) and with only a limited number of human scores. The data we have are in steps of $\frac{1}{3}$ which makes the data discrete, however, if we had more samples we could get a smoother average which could possibly bring us closer to the normal distribution CDF. This aligns with what we found in our KS test that the average human scores are not drawn from a normal distribution.

5 DISCUSSION AND LIMITATIONS

The limitation of this project mainly lies in the data collection process. After reviewing the Twitter data, we found that a majority of the Twitter dataset is posts that have very strong political positions and offensive language. Thus, it could be a biased dataset, and our results might be less representative in terms of sentiment analysis performance on a general text corpus. Furthermore, when collecting humans' opinions on the rating of the sentences, we only collect ratings of 250 sentences from three participants. A more comprehensive dataset of human-labeled sentiment scores on a continuous scale could reduce the bias introduced in the labeling process.

As for future directions, in addition to creating a more comprehensive dataset for the evaluation of sentiment analysis, we plan to test how well the conventional models generalize to a completely different dataset. This could be fairer to LLMs since they are trained on a more diverse language corpus and thus may perform worse than the conventional models, which are trained on the same Twitter dataset. We could also fine-tune the LLMs on the sentiment analysis task, which could enhance the LLMs sentiment alignment with humans.

REFERENCES

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems* 32, 6 (2017), 74–80.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[8] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[9] Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294* (2023).

[10] OpenAI. 2023. ChatGPT: Optimizing Language Models for Dialogue. *OpenAI* (2023). https://openai.com/chatgpt.

[11] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[12] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[14] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. 502–518.

[15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903* (2022).

[17] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).

[18] Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*. PMLR, 41092–41110.

[19] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1253.

[20] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005* (2023).

A APPENDIX

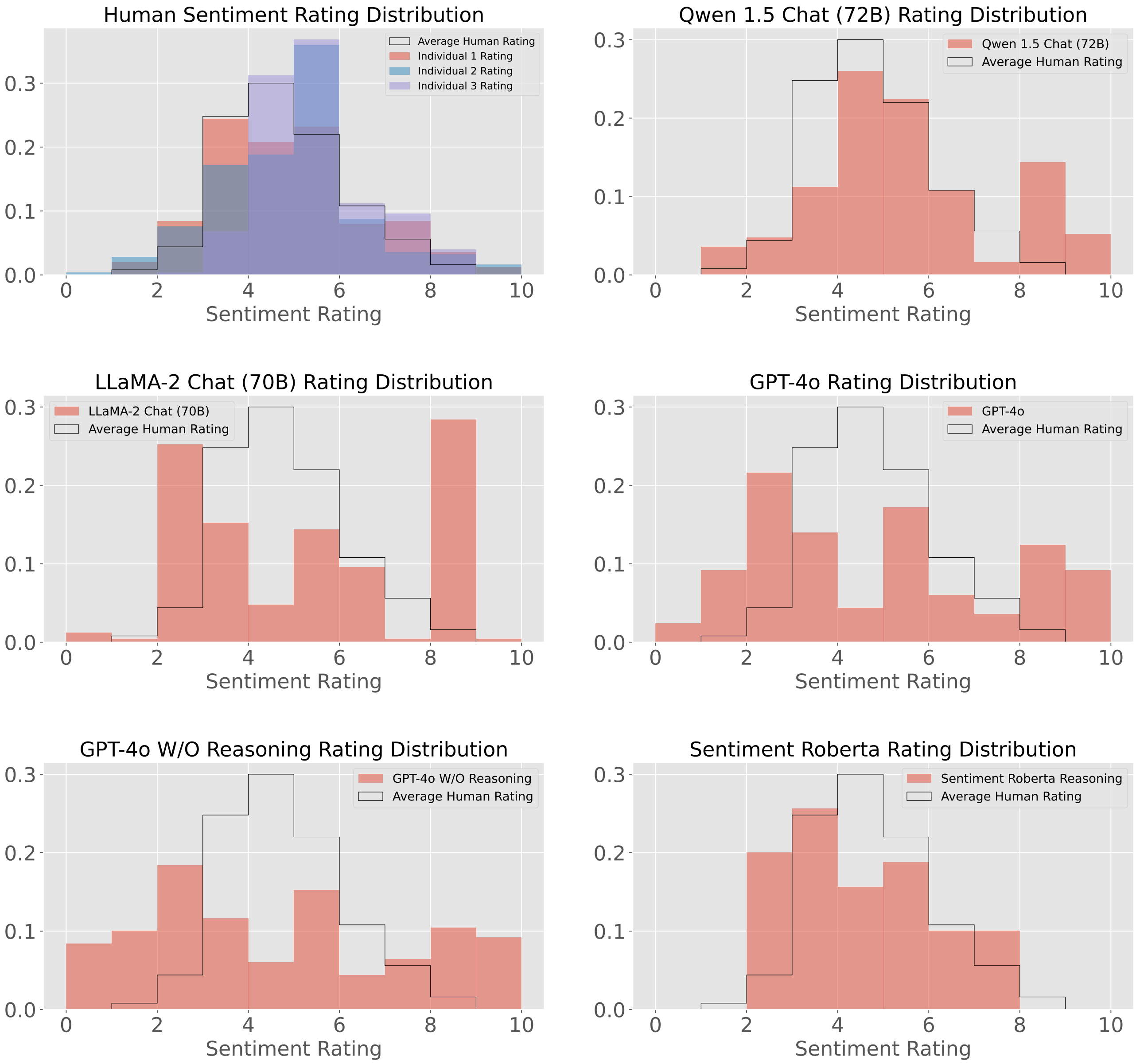


Figure 6: Histograms displaying the distribution of sentiment scoring from select models, both LLMs and a conventional model. The distribution of the average human scoring is also shown, drawn from the three project members.