The limits of excellence: Assessing fine-tuned ChatGPT's efficacy in stock price forecasting

Yiyang Huang[1,5,9], **Xiang Liu**[2,6], **Naichuan Zhang**[3,7], **Tianshu Zhao**[4,8]

[1]School of Electrical Engineering and Artificial Intelligence, Xiamen University Malaysia, Selangor, 43900, Malaysia
[2]Information Sciences and Technology, Penn State University, Pennsylvania, 16801, United State
[3]School of Ulster, Shannxi University of Science & Technology, Xian,710016,China
[4]School of Software, Jilin University, Changchun, 130015, China

[5]AIT2009365@xmu.edu.my
[6]xfl5249@psu.edu
[7]FiTdF628@Gmail.com
[8]zhaots5520@mails.jlu.edu.cn
[9]corresponding author

Abstract. In this study, we explore the ability of ChatGPT to predict stock market trends based on stock news headlines and real stock market data. In order to evaluate the performance of the fine-tuned model, we first obtain the prediction results of GPT-3.5 Turbo on specific stock's future trends as a comparison. We fine-tuned GPT-3.5 Turbo and conducted related training, testing and result evaluation. The experiments implemented on the two datasets Bigdata2022 and Cikm illustrate that fine-tuning can help the model to produce expected structured output according to user requirements, based on its more sophisticated understanding of the text and data in this field. However, although the model's performance is improved significantly, GPT-3.5 Turbo does not demonstrate better performance compared to other traditional large language models in terms of integrating time series data and news headline data for stock forecasting. The fine-tuned ChatGPT model is expected to achieve excellent results in the stock market forecasting tasks through more in-depth research and become one of the mainstream research models in this field.

Keywords: Large Language Models, GPT-3.5 Turbo, Fine-Tuning, Return Predictability, Textual Analysis

1. Introduction

In recent years, the application of generative artificial intelligence and large-scale language modeling (LLM) has received a lot of attention with a proliferation of related research on the potential of application in different fields. Among the various research orientation, the study of LLM in financial economics especially in stock returns prediction, is still immature, and the models have not been specifically trained for this domain. But the nature of their training process which is based on real large text data enables them to understand the contextual information in natural language better, so it is still believed that LLM is promising for predicting stock returns.

Predicting stock returns through news and algorithms is nothing new, there has been sufficient studies to prove this [1] [2]. Based on this, our goal is to evaluate whether an LLM that has not been trained to

predict returns can acquire the ability to predict stock returns by using news and training the algorithm. We use a stock dataset which is derived from real U.S. stock markets with large trading volumes [3]. We perform data preprocessing operations on the data and divide the data set into three parts: training, validation and testing. In keeping with usual stock research, the three sections are arranged chronologically. The dataset contains many details such as a stock's opening, high, low, closing and adjusted closing prices, as well as average prices over different time lengths.

Through experiments on two datasets Bigdata22 and Cikm18, we explore the application potential of ChatGPT 3.5 Turbo and its fine-tuned version in predicting stock trends based on news headlines, and evaluate the model effect using a specific evaluation method for LLM[4]. The results show that in the application of stock prediction by combining time series and news headlines, GPT3.5-Turbo's performance after fine-tuning has improved compared with before fine-tuning, thanks to its ability to structure output and enhance fine-tuning Data domain understanding. However, its performance is still slightly inferior compared to other LLMs.

2. Background

LLM and Stock Price Predictions

Large language models (LLM) is an important milestone in the field of NLP. LLM is an artificial intelligence model based on deep neural network, this feature equips the models with excellent natural language understanding capabilities. The uniqueness of LLM lies in its pre-training ability: it can learn language patterns, grammar, semantics and other knowledge through large-scale text, and then adapt to specific tasks through fine-tuning process. Some mainstream LLM models have achieved remarkable results, including BERT, GPT-3, and GPT-3.5 Turbo used in this study.

BERT (Bidirectional Encoder Representations from Transformers):
BERT has an innovation lies in bidirectional pre-training, which can better understand contextual information. BERT has achieved significant performance improvements in various NLP tasks, providing strong support for text understanding and classification tasks.

GPT-3 (Generative Pre-trained Transformer 3):
GPT-3 is one of the largest LLMs to date With 175 billion parameters. It performs well in multiple NLP tasks, including text generation, sentiment analysis, text classification, etc. The power of GPT-3 lies in its ability to generate high-quality natural language text, which has broad application potential.

GPT-3.5 Turbo:
GPT-3.5 Turbo is an enhanced version of GPT-3 with excellent text generation and understanding capabilities. The model has hundreds of billions of parameters, supports multi-lingual tasks, and performs well in multiple fields, including text summarization, dialogue generation, and stock market prediction, which is the focus of this study.

Stock market prediction has always been a popular topic in the financial field. Most traditional stock forecasting methods mainly rely on tools such as technical analysis, fundamental analysis, and market sentiment analysis, but these methods are faced with the limitation problem since they cannot fully learn the mechanisms of how the large amounts of information bring impact to the stock market, while this ability is the advantage of LLM, which can extract information and learn the patterns from complex text data.

Relevant research has proven that LLM performs well in natural language processing tasks such as sentiment analysis, text classification, and named entity recognition. However, the application of LLM in the field of stock market prediction is still in a relatively early stage. By using LLM for predictive analysis after combining stock market data and news text data, we can predict the trends, volatility and future returns of stock market with better understanding of implicit relationship between stock price

and text associated with it, resulting in better investment decisions and risk management. Therefore, this study will focus on the LLM technology GPT-3.5 Turbo to explore its application potential and limitations in stock market prediction.

We received certain inspiration after previous investigation in this domain. The comprehensive zero-shot analysis "The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over MultiModal Stock Movement Prediction Challenges" [3] explores ChatGPT's capabilities in predicting stock movements using tweets and historical stock price datasets. Their study underlines the need for domain-specific training or fine-tuning, especially when applying models like ChatGPT for financial forecasting, given its underperformance in comparison to both contemporary and traditional methods.Further, "PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance" [5] introduces a comprehensive framework. This includes the first financial LLM, FinMA, complemented with instruction data and an evaluation benchmark. The research underscores the potential of financial LLMs across a spectrum of financial tasks. The research "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models" [6] delves into the potential of ChatGPT and other LLMs in predicting stock market returns based on news headlines. Their findings advocate the superior performance of advanced models like ChatGPT-4 in return predictability.

3. Method

3.1. Data

In previous research, they construct the dataset by first obtaining daily stock return data for all U.S. common stocks from the CRSP database and then collecting news headlines related to these stocks from mainstream news outlets and newswires [6]. Since the training data of ChatGPT ends in September 2021, the sample period selected ranges from October 2021 to December 2022, thereby ensuring that the evaluation is based on data information that does not exist in the ChatGPT model training data, can more accurately assess its predictive ability on new data. ChatGPT evaluates the headline's impact on the company's stock price, expressing the assessment as good, bad, or neutral. These evaluations are converted into numerical scores that can be used to predict stock returns for the next trading day.

In our study on stock movement prediction, two primary benchmark datasets were utilized: BIGDATA22 [7], and CIKM18 [8]. These datasets consist of stocks from the U.S. stock markets that have substantial trading volume and were compiled by Xie, Q. and their team members [3]. They featured both historical price details and relevant tweet data. To extract and align important features from these datasets, the data underwent a preprocessing phase, ensuring all datasets had a unified representation for our analysis. The dataset is divided into three parts: training, validation, and test. These parts are arranged by time, just like in usual stock studies. The datasets have many details like the opening, high, low, closing, and adjusted close prices of stocks. They also have average prices over periods like 5, 10, 15, 20, 25, and 30 days.

3.2. Model — GPT-3.5 Turbo

GPT-3.5 Turbo, an advanced derivative of the GPT-3.5 family, is optimized primarily for chat applications but also demonstrates versatility across various tasks. This model, noted as the most proficient within its lineage, excels in understanding and generating both natural language and code, making it suitable for diverse applications beyond chat functionalities [9]. A key feature of GPT-3.5 Turbo is its fine-tuning capability via OpenAI's API, which enhances performance on specific applications. While the pre-trained model is competent through prompt engineering, fine-tuning elevates its performance, occasionally surpassing GPT-4 in narrow tasks as evidenced by preliminary tests. In the domain of stock price prediction, the attributes of GPT-3.5 Turbo may present substantial advantages. Its adept understanding of language and code, combined with the ability to fine-tune via OpenAI's API, provides a potential platform for delving into complex financial data and attempting to formulate grounded predictions. The fine-tuning capability not only offers a means to enhance the model's performance but also aligns it more closely with the nuanced needs of developers, thereby customizing

the AI model's capabilities to the task at hand. In essence, through meticulous fine-tuning, GPT-3.5 Turbo can be adeptly calibrated to meet the demands of stock price prediction projects, marking a significant stride towards more accurate and domain-specific predictive analysis. Its proficiency in comprehending language and code, amalgamated with fine-tuning capabilities, underscores its significance and utility in financial analytics and stock price forecasting, providing favorable conditions for further exploration in this domain.

3.3. Prompt

In the endeavor to tailor machine learning models for specialized tasks, the employment of prompts has surfaced as a crucial technique, particularly in the realm of fine-tuning models like GPT-3.5 Turbo for stock price prediction. Prompts act as a mechanism to instruct the model on the task at hand, steering its response generation towards the desired outcome. They function as a conduit between the raw data and the model, offering a structured avenue for data input and soliciting specific types of analysis or predictions. Notably, the importance of judicious prompt design escalates, especially when training models with an extensive number of parameters such as GPT [10]. The architecture of GPT-3.5 Turbo inherently facilitates the use of prompts owing to its transformer-based design, which is amenable to comprehending and generating responses predicated on contextual cues provided in the prompts. The prompt employed in this research, for instance, is structured as follows:

```
Scrutinize the data and tweets to envisage if the closing price
of $mdt will swell or contract at 2020-04-01.  Please affirm either
Rise or Fall.
[Time-series_stock_price]
[News_headline]
Answer:
```

The employment of prompts serves multifaceted purposes in fine-tuning the GPT-3.5 Turbo for stock price prediction. Primarily, the prompt steers the model towards the task of analyzing stock price movement predicated on the furnished data and social media discourse. It encapsulates the quintessence of the task: to ascertain whether the closing price of a specific stock will ascend or descend on a designated date. Secondly, the structured demeanor of the prompt ensures the model is furnished with all requisite data to render an informed prediction, encompassing historical stock price data and pertinent social media content. Moreover, the utilization of a meticulously crafted prompt can markedly augment the model's cognizance of the task, potentially enhancing the accuracy and relevance of its predictions [11]. The organized nature of the prompt employed in this study facilitates the model's understanding of the underlying data and the prediction task. It delineates a clear indication of the desired output format, which is indispensable for evaluating the model's performance in a consistent and standardized fashion. Furthermore, prompts contribute to a more structured and nuanced output, bolstering various software engineering tasks in the context of LLMs (Large Language Models) [12].

The utilization of prompts is a cardinal aspect of fine-tuning GPT-3.5 Turbo for stock price prediction. It proffers a structured, lucid, and effective conduit to guide the model towards engendering accurate and pertinent predictions predicated on the provided data. This methodology, orchestrated around a well-conceived prompt, harbors promise for augmenting the model's performance in stock price prediction endeavors, and potentially, in other domain-specific prediction tasks as well.

4. Implement

4.1. Fine-Tuning

4.1.1. Data Preprocessing Given the incongruity between the format of the original dataset's prompt and the standard prompt format stipulated by OpenAI, an initial step entailed the modification of the training and validation sets to align with the format mandated by OpenAI documentation. Subsequently, the modified training and validation sets were encapsulated into JSONL files which were duly uploaded.

4.1.2. Fine-Tuning Setup Utilizing the OpenAI API, a fine-tuning task was instantiated. The model of choice was GPT-3.5 Turbo, and the uploaded training and validation sets were selected for the task. The hyperparameters were meticulously configured with an epoch value of 2, as illustrated in the code below:

```
openai.FineTuningJob.create(
training_file="TRAINING_FILE",
validation_file="VALIDATION_FILE",
model="gpt-3.5-turbo",
hyperparameters="n_epochs":2)
```

4.1.3. Training Result Visualization OpenAI provided a method to keep the training results throughout the training process. This feature made it possible to record crucial parameters, such as training accuracy and loss, at each epoch. Following training, the accumulated results were saved and put via visualization, providing a clear picture of the model's performance and convergence tendencies.

In our maiden endeavor of fine-tuning the GPT-3.5 Turbo model, the focus was primarily on the Bigdata22 dataset. In Figure 1, The preliminary observation showcased an elevated training loss of 4.0, which was anticipated due to the unoptimized nature of the model concerning this specific dataset. However, a swift descent in the loss was recorded, plummeting to around 1.5 post approximately 200 steps. Concurrently, an ascension in accuracy was noted, catapulting from around 0.5 to approximately 0.85, indicating a rapid learning pace of the model on the Bigdata dataset at the inception of fine-tuning. Post this initial rapid learning phase, the loss trajectory exhibited a plateau, with accuracy oscillating around 0.85. This plateau potentially underscores challenges the model might be encountering with certain data points or features, reflecting a zone of stabilized learning yet with room for refinement.

The narrative took an intriguing turn when the model, now fine-tuned on Bigdata dataset, was subjected to a subsequent fine-tuning on the Cikm dataset. Figure 2 depicts the training results during this phase. The results exhibit a stark contrast with the initial training loss being recorded around 0.4, substantially lower than the initial loss during the first fine-tuning phase. This observation suggests that the training on Bigdata22 endowed the model with a certain degree of adaptability and generality, despite Cikm being a disparate dataset. The loss diminished to around 0.1 at about 250 steps, while the accuracy soared swiftly to nearly 0.9, maintaining a steadiness throughout the majority of the training period, with only minor fluctuations observed towards the latter phase. This scenario accentuates that the preliminary fine-tuning on Bigdata significantly bolstered the model's adaptability and stability when confronted with the Cikm dataset.

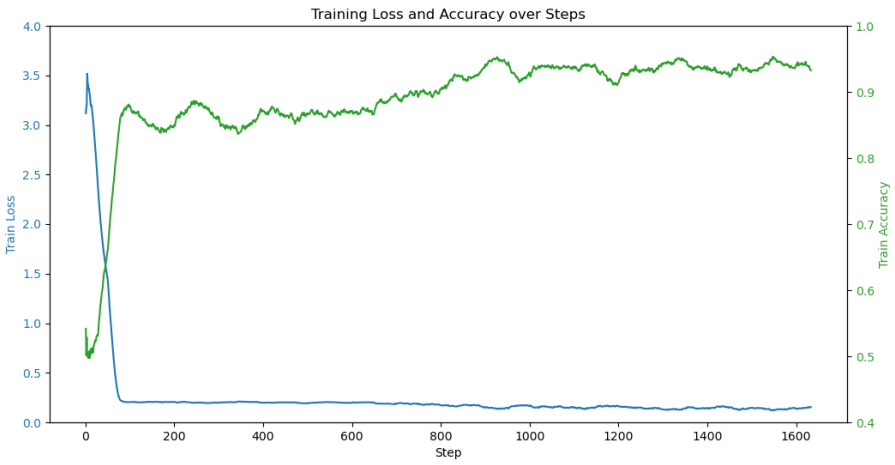

Figure 1: Training Result of GPT-3.5 Turbo (Bigdata22)

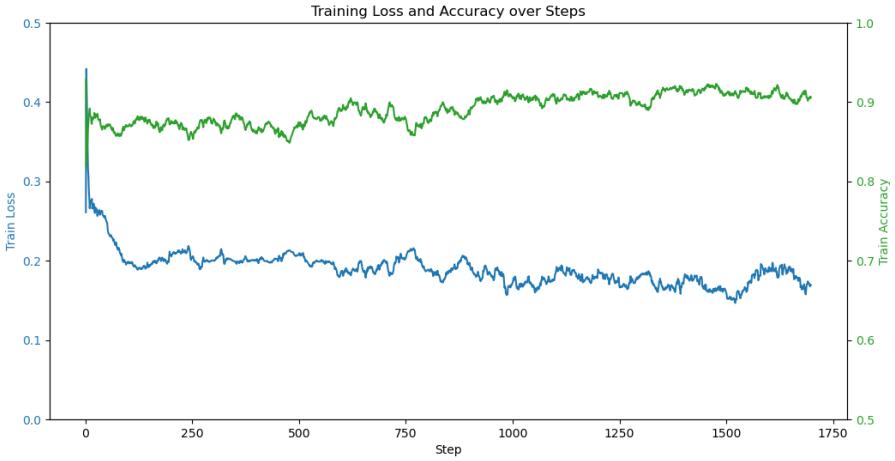

Figure 2: Training Result of Fine-tuned GPT-3.5 Turbo (Cikm)

The GPT-3.5 Turbo model's first fine-tuning on the Bigdata dataset established a solid foundation that showed in excellent performance throughout the model's subsequent fine-tuning on the Cikm dataset. This emphasizes the critical importance of a model's initial fine-tuning on a substantially larger or more complicated dataset, which may provide the model greater adaptability and improve its performance on subsequent tasks or datasets.

4.2. Evaluation

4.2.1. Metrics In line with prior methodologies, we utilize evaluation metrics such as Accuracy (ACC), F1 Score, Recall, and Bootstrap Standard Deviation (Bootstrap Std) to evaluate the capability of ChatGPT and other benchmark models in forecasting stock movements based on news headlines. These chosen metrics offer a holistic view, enabling us to determine the efficacy of predictions in the context of stock trends, particularly factoring in the distribution of up trending and down trending instances [13].

Accuracy(ACC): The proportion of correct predictions among the total number of predictions. It is calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Where TP represents the count of true positives, TN signifies true negatives, FP is indicative of false positives, and FN captures false negatives. As a prevalent metric in classification endeavors, Accuracy furnishes a direct insight into the model's predictive prowess.

Recall: Recall is a key indicator to evaluate the performance of a classification model. Its main purpose is to measure the model's ability to correctly identify positive examples. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

The value range of recall is [0, 1]. It measures what proportion of all positive examples the model can correctly detect.

F1-Score: It is an indicator used in statistics to measure the accuracy of a binary classification model. It considers both the precision and recall of the classification model. The F1 score can be viewed as a harmonic average of model precision and recall. It is calculated as follows:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

The value range of F1 score is [0,1], where 1 represents the best performance of the model and 0 represents the worst performance.

4.2.2. Output Analysis In our study, we conducted an in-depth evaluation of OpenAI's original ChatGPT-3.5 Turbo model and our fine-tuned version. The experimental results clearly showcased significant differences in the outputs of the two.The example of ChatGPT3.5 Turbo with irrelevant prediction in Bigdata22 dataset is shown in Table 1, and the result of fine-tuned model in Table 2.

The unmodified ChatGPT-3.5 Turbo, when responding to prompts, tends to offer comprehensive descriptions and analyses rather than merely rendering succinct predictions like "Rise" or "Fall". For instance, when queried about the stock price trend, the model might elaborate, "Based on the provided data and tweets, determining with certainty whether the closing price of $hd will surge or decline on 2020-12-16 is challenging. However, the closing prices appear to fluctuate across the given dates. Thus, it's more probable that the closing price will persist in its fluctuation rather than exhibiting a distinct trend of either ascending or descending." This level of detailed response furnishes users with enriched context and a deeper interpretation of the data, facilitating an understanding of the rationale behind predictions. As depicted in Table 1, even though the model furnishes such thorough descriptions, its predictions on the Bigdata22 dataset still diverge from the actual outcomes. Yet, despite this depth of analysis, the original model's performance in prediction accuracy is not as satisfactory as expected. Moreover, it sometimes fails to render clear analyses on certain data points. Notably, both GPT3.5 Turbo and GPT4 demonstrate relatively low predictive accuracy [3].

However, when we evaluated the fine-tuned model on the same dataset, all the outputs were definitive predictions, such as "Rise" or "Fall". Despite this streamlined output approach has not a detailed context analysis of user input, the performance of the model is greatly improved compared to the original model. As we proved before, the original model always generates mumble responses to avoid making explicit judgments. So that it influences the users' experience. But this issue is absent in the fine-tuned version.

For instance, as illustrated in Table 2, the fine-tuned model definitively predicted a "Fall" for the closing price on 2020-12-16 in the Bigdata22 dataset and this prediction aligned with the actual outcome. The uncertainty and ambiguity of the original model contrast sharply with the results of our fine-tuned model. It further proves that after human training and fine-tuning, ChatGPT3.5 Turbo can show high accuracy and purpose in this task.

Type	Content
Prompt	Context:date,open,high,low,close,adj-close,inc-5,inc-10,inc-15,inc-20,inc-25,inc-30 2020-12-02,1.1,1.1,-0.6,-2.0,-1.5,1.0,0.2,0.6,1.1,0.8,1.1 2020-12-03,0.9,1.4,-0.4,-1.1,-1.1,1.8,1.3,1.6,2.0,1.9,2.1 2020-12-04,1.3,1.8,-0.4,-1.5,-1.5,2.6,2.7,2.9,3.2,3.4,3.5 2020-12-07,0.3,0.7,-0.6,-0.5,-0.5,2.1,3.0,3.1,3.4,3.9,3.8 2020-12-08,0.0,0.3,-1.1,-0.4,-0.4,1.5,3.1,3.0,3.6,4.2,4.0 2020-12-09,-0.4,0.1,-1.3,1.5,1.5,-0.4,1.3,1.4,1.9,2.5,2.3 2020-12-10,-0.3,0.5,-1.0,-0.3,-0.3,-0.4,1.3,1.6,2.0,2.6,2.6 2020-12-11,-0.4,0.3,-0.7,-0.1,-0.1,-0.3,1.1,1.6,1.9,2.4,2.6 2020-12-14,0.2,1.1,-0.1,0.4,0.4,-0.4,0.3,1.1,1.3,1.7,2.2 2020-12-15,-0.2,0.3,-0.7,0.9,0.9,-0.9,-0.9,0.2,0.2,0.8,1.3 Contemplate the data and tweets to guess whether the closing price of $MMM will surge or decline at 2020-12-16. Please declare either Rise or Fall.
Response	Based on the given data and tweets, it is difficult to determine with certainty whether the closing price of $hd will surge or decline on 2020-12-16. However, we can observe that the closing prices have been fluctuating throughout the given dates. Therefore, it is more likely that the closing price will continue to fluctuate rather than showing a clear trend of either rising or falling.
Prediction	Null
Truth	Rise

Table 1: ChatGPT3.5 Turbo in Bigdata22 dataset

Type	Content
Prompt	Context:date,open,high,low,close,adj-close,inc-5,inc-10,inc-15,inc-20,inc-25,inc-30 2020-12-02,1.1,1.1,-0.6,-2.0,-1.5,1.0,0.2,0.6,1.1,0.8,1.1 2020-12-03,0.9,1.4,-0.4,-1.1,-1.1,1.8,1.3,1.6,2.0,1.9,2.1 2020-12-04,1.3,1.8,-0.4,-1.5,-1.5,2.6,2.7,2.9,3.2,3.4,3.5 2020-12-07,0.3,0.7,-0.6,-0.5,-0.5,2.1,3.0,3.1,3.4,3.9,3.8 2020-12-08,0.0,0.3,-1.1,-0.4,-0.4,1.5,3.1,3.0,3.6,4.2,4.0 2020-12-09,-0.4,0.1,-1.3,1.5,1.5,-0.4,1.3,1.4,1.9,2.5,2.3 2020-12-10,-0.3,0.5,-1.0,-0.3,-0.3,-0.4,1.3,1.6,2.0,2.6,2.6 2020-12-11,-0.4,0.3,-0.7,-0.1,-0.1,-0.3,1.1,1.6,1.9,2.4,2.6 2020-12-14,0.2,1.1,-0.1,0.4,0.4,-0.4,0.3,1.1,1.3,1.7,2.2 2020-12-15,-0.2,0.3,-0.7,0.9,0.9,-0.9,-0.9,0.2,0.2,0.8,1.3 Contemplate the data and tweets to guess whether the closing price of $MMM will surge or decline at 2020-12-16. Please declare either Rise or Fall.
Response	Fall
Prediction	Fall
Truth	Fall

Table 2: Fine-tuned Model in Bigdata22 dataset

4.2.3. Main Results The evaluation of experimental results is based on OpenAI evals, which is a model evaluation tool proposed by OpenAI to compare various aspects of performance of different LLMs [4]. We compared different models for stock price prediction using F1-score, accuracy (Acc), recall metrics on the two datasets: Bigdata22 and Cikm18.

	F1-Score	ACC	Recall
ChatGPT3.5 Turbo	0.4310	0.1630	0.2747
Fine-tuned GPT3.5	0.8336	0.5149	0.7147

Table 3: A comparison of different models on dataset Bigdata22

	F1-Score	ACC	Recall
ChatGPT3.5 Turbo	0.4573	0.1601	0.2965
Fine-tuned GPT3.5	0.6488	0.4987	0.4802

Table 4: A comparison of different models on dataset CIKM18

Table 3 presents our core experimental results on the Bigdata22 dataset. It is evident from the outcomes that the fine-tuned model demonstrates significant superiority across various metrics in comparison to the original version. Specifically, on the Bigdata22 dataset, the fine-tuned model achieved an accuracy of 0.5149, substantially surpassing the original model's 0.1630. This marked improvement reaffirms the enhanced prediction precision of the fine-tuned model on this dataset. Besides, its recall rate increased from 0.2747 to 0.7147, further proving the fine-tuned model's improved efficacy in identifying correct output samples. Concurrently, there was a notable enhancement in the F1 score.

Regarding the CIKM18 dataset, the fine-tuned model also displayed a positive trajectory as shown in Table 4. However, because the CIKM18 dataset contains more complex user input, such as longer news titles and more text. So, the performance of the fine-tuned model on this dataset is slightly worse than that on Bigdata22.

To sum up, whether on the Bigdata22 or CIKM18 dataset, the fine-tuned GPT3.5 Turbo showed its high performance in F1-Score, accuracy, and recall rate. However, there are differences in the performance improvement between the two datasets, which may be related to the inherent characteristics, data distribution, and size of the datasets.

5. Conclusion

This study delves deeply into the potential applications of ChatGPT 3.5 Turbo and its fine-tuned version in predicting stock trends based on news headlines. To conduct experiments, we used two different datasets to compare between the original During the experiments, we use two different datasets to compare the original model and its fine-tuned version. After our evaluation, the fine-tuned ChatGPT 3.5 Turbo showed a significant increase in accuracy. This is evident in two main aspects: One of that aspect is the fine-tuned model could produce more structured outputs, reducing ambiguous and non-targeted feedback, and better aligning with user output expectations; another aspect is the model's insight and understanding of the financial stock domain saw substantial enhancements.

Despite the fine-tuning, ChatGPT 3.5 Turbo, when integrating time-series data with news headlines for stock predictions, still exists noticeable shortcomings in comparison to traditional prediction models. This shows that, although there have been some achievements, ChatGPT still has a lot of room for improvement and prospects in the field of stock prediction.

Our conclusion highlights the potential and advantages of ChatGPT 3.5 Turbo in stock market prediction, especially that fine-tuning can improve its ability to generate structured output and improve domain understanding, making it better suited to tasks in specific fields. However, models still face significant challenges when dealing with time series data and more complex market scenarios. Our study provides valuable research directions for the problem of how to better utilize large-scale language models

in stock market prediction.

6. Future Work

During the course of this study, we fine-tuned GPT-3.5 Turbo and obtained preliminary results. However, the range of potential directions for further exploration and optimization of the application of this model is wide:

Model structure and rich domain knowledge: Although fine-tuning significantly enhances GPT-3.5 Turbo's output and understanding capabilities in specific data domains, there is room for further improvement in prediction accuracy. To achieve this goal, future research can further optimize the model's architecture to better meet the unique requirements of the financial field. Integrating more domain-specific data, such as historical stock prices and market indicators, has the potential to improve model performance in stock prediction tasks.

Fine-tuning methods and model interpretability: Despite our efforts to fine-tune the model and observe progress, there remains untapped potential for further refinement. Given the complexity of financial forecasting, we plan to explore fine-tuning methods tailored to the specific needs of this area. In addition, improving the interpretability of the model is a focus. Enabling GPT-3.5 Turbo to provide detailed reasons during the prediction process, allowing users to have a deeper understanding of their decisions, is also a problem that needs to be improved.

While our initial results are encouraging, it is clear that GPT-3.5 Turbo has significant untapped potential and numerous opportunities for further exploration and optimization in the field of financial prediction.

Acknowledgement

Yiyang Huang, Xiang Liu, Naichuan Zhang and Tianshu Zhao contributed equally to this work and should be considered co-first authors.

References

[1] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168, 2007.

[2] Anastassia Fedyk and James Hodson. When can the market identify old news? *Journal of Financial Economics*, 149(1):92–113, 2023.

[3] Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. *arXiv preprint arXiv:2304.05351*, 2023.

[4] Github Community. Evals is a framework for evaluating llms and llm systems, and an open-source registry of benchmarks., 2023.

[5] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

[6] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

[7] jiminHuang. Datasets:chancefocus/flare-sm-bigdata, 2022.

[8] jiminHuang. Datasets:chancefocus/flare-sm-cikm, 2022.

[9] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

[10] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[11] Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*, 2022.

[12] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

[13] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.