FinLLM-B: When Large Language Models Meet Financial Breakout Trading

Kang Zhang[1,2], Osamu Yoshie[2], Lichao Sun[3], Weiran Huang[1,4,*]

[1]MIFA Lab, Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China
[2]Waseda University, Tokyo, Japan [3]Lehigh University, Bethlehem, PA, USA
[4]State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China

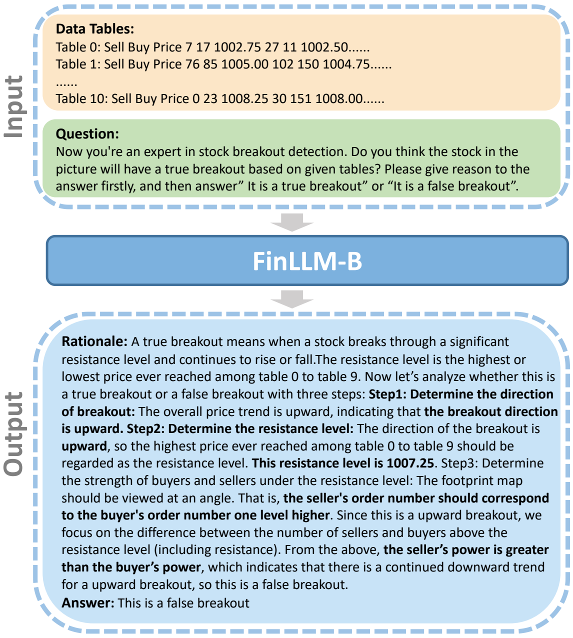
zhangkang@toki.waseda.jp, yoshie@waseda.jp, lis221@lehigh.edu, weiran.huang@outlook.com

Abstract

Trading range breakout is a key method in the technical analysis of financial trading, widely employed by traders in financial markets such as stocks, futures, and foreign exchange. However, distinguishing between true and false breakout and providing the correct rationale cause significant challenges to investors. Traditional quantitative methods require large amounts of data and cannot directly present the reasoning process, making them less than perfect in this field. Recently, large language models have achieved success in various downstream applications, but their effectiveness in the domain of financial breakout detection has been subpar. The reason is that the unique data and specific knowledge are required in breakout detection. To address these issues, we created the first financial breakout dataset and introduce FinLLM-B, the premier large language model for financial breakout detection, which enhances the effectiveness of breakout trading strategies. Furthermore, we have developed a novel framework for large language models, namely multi-stage structure, effectively reducing mistakes in downstream applications. Experimental results indicate that compared to GPT-3.5, FinLLM-B improves the average accuracy of answers and rational by 49.97%, with the multi-stage structure contributing 9.72% to the improvement. Additionally, it outperforms ChatGPT-4 by 42.38%.

1 Introduction

Fundamental and technical analysis are the primary methods in financial investment. Given the limitations of the efficient market hypothesis in real financial markets (Ball, 2009; Malkiel, 2003; Stout, 2002), the significance of technical analysis is recognized (Blume et al., 1994; Taylor and Allen, 1992; Lo et al., 2000; Knight, 2010). Trading range breakouts, a key aspect of technical analysis, have been shown to be highly effective (Zhu et al., 2015; Lubnau and Todorova, 2014). However, breakouts are often accompanied by false signals, which makes detecting false breakouts an important issue of breakout trading strategy (Zhang et al., 2020; Elder, 2002). Traditional quantitative methods (Han and Fu, 2023; John and Latha, 2023; Zhang et al., 2023a; Kim and Kim, 2019) struggle with breakout detection due to limitations in dataset accessibility and report readability. For dataset accessibility, breakout detection requires footprint data, which is not readily available in mainstream datasets, hindering model training. For report readability, the finance sector demands high model explainability to ensure transparent decision-making (Laux et al., 2024; Ben David et al., 2021; Fritz-Morgenthal et al., 2022). Addressing these challenges is crucial for improving breakout detection methods.



Data Tables:
Table 0: Sell Buy Price 7 17 1002.75 27 11 1002.50......
Table 1: Sell Buy Price 76 85 1005.00 102 150 1004.75......
......
Table 10: Sell Buy Price 0 23 1008.25 30 151 1008.00......

Question:
Now you're an expert in stock breakout detection. Do you think the stock in the picture will have a true breakout based on given tables? Please give reason to the answer firstly, and then answer” It is a true breakout” or “It is a false breakout”.

Rationale: A true breakout means when a stock breaks through a significant resistance level and continues to rise or fall.The resistance level is the highest or lowest price ever reached among table 0 to table 9. Now let’s analyze whether this is a true breakout or a false breakout with three steps: **Step1: Determine the direction of breakout:** The overall price trend is upward, indicating that **the breakout direction is upward. Step2: Determine the resistance level:** The direction of the breakout is **upward**, so the highest price ever reached among table 0 to table 9 should be regarded as the resistance level. **This resistance level is 1007.25**. Step3: Determine the strength of buyers and sellers under the resistance level: The footprint map should be viewed at an angle. That is, **the seller's order number should correspond to the buyer's order number one level higher**. Since this is a upward breakout, we focus on the difference between the number of sellers and buyers above the resistance level (including resistance). From the above, **the seller’s power is greater than the buyer’s power**, which indicates that there is a continued downward trend for a upward breakout, so this is a false breakout.
Answer: This is a false breakout

Figure 1: A sample of FinLLM-B task.

Large language models (LLMs) have shown promise in fine-tuning with limited data (Brown

[*] Correspondence to Weiran Huang. A previous version of this work is available at arXiv:2402.07536.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 349–357
April 30, 2025 ©2025 Association for Computational Linguistics

et al., 2020; Gao et al., 2020) and generating comprehensive reports with rationale. These characteristics make LLMs strong candidates for breakout detection. However, three challenges remain. Firstly, LLMs lack domain knowledge, as observed in our experiments with GPT-3.5 and GPT-4, which struggled with breakout detection queries due to insufficient specialized datasets. Secondly, LLMs often produce outputs with mistakes (McIntosh et al., 2023; Lee, 2023; Zhang et al., 2023b), including incorrect resistance levels and trend analysis. Thirdly, LLMs exhibit output inconsistency (Chang et al., 2024; Tan et al., 2023), which can significantly impact model performance in financial domain.

In this work, we introduce FinLLM-B, a LLM for financial breakout detection as shown in Figure 1. FinLLM-B supplements the foundational knowledge of GPT-3.5 in breakout detection and employs a multi-stage framework to mitigate errors and instability. This framework segments the rationale, allowing FinLLM-B to focus on subtasks, improving both accuracy and stability. Our experiments show that FinLLM-B outperforms GPT-3.5, achieving a 49.97% improvement.

Our contributions can be summarized as follows: 1) We introduce FinLLM-B, the first large language model for financial breakout detection, which demonstrates domain knowledge and helps improve the reliability of breakout trading strategies. 2) Financial breakout dataset. We create the first dataset for financial breakouts, providing a valuable resource for future research in this area. 3) Multi-stage structure. We propose a multi-stage structure that segments the rationale, effectively reducing errors and enhancing stability for large language models in downstream tasks.

2 Related Work

Trading Range Breakout. Technical analysis focuses on predicting financial market movements based on historical chart data (Murphy, 1999), demonstrating its profitability (Taylor and Allen, 1992; Lo et al., 2000). A key method within technical analysis is the trading range breakout (Raj and Thurston, 1996; Lento et al., 2007; Bessembinder and Chan, 1995), which suggests that a price struggle occurs between buyers and sellers at resistance levels. Once the price surpasses this resistance level, it forms a strong support, preventing a short-term price reversal (Brooks, 2011; Chordia et al., 2002; Gosnell et al., 1996).

Large Language Models. Large language models (LLMs) have shown success across various applications (Wu et al., 2023; Li et al., 2023; Luo et al., 2022; Bi et al., 2023; Kraljevic et al., 2021; Sarrion, 2023; Liu et al., 2023, 2021; Li et al., 2024). A challenge of applications is generation of incorrect answers. One solution related to this study is chain-of-thought (CoT) (Wei et al., 2022) which prompts LLMs to reason before providing answers. Pioneering works involved manually designing examples to teach models reasoning, enabling more accurate responses (Wei et al., 2022). Subsequent research introduced approaches like zero-shot-CoT (Kojima et al., 2022) and auto-CoT (Zhang et al., 2022), though CoT does not fully eliminate incorrect outputs, and researchers have explored incorporating new modalities (Zhang et al., 2023c; Lu et al., 2022).

3 Problem Formulation

Financial breakout detection is an important problem in the field of breakout trading. It determines whether a financial product is undergoing a true or false breakout, with true breakouts identified based on the order flow rule (Valtos, 2015). This study focuses on training a large language model to generate financial breakout detection reports with accurate rationales using processed data tables.

Time scale variability affects resistance levels and breakout authenticity, requiring a clear definition of the resistance level and true breakouts. The resistance level is defined as the highest or lowest price in the ten time ticks before the breakout (Brooks, 2011; Valtos, 2015). A true breakout occurs when the closing price remains beyond the resistance level for two consecutive time units.

The primary input is a data table as shown in Figure 1 derived from footprint charts. These charts capture detailed price information within each time unit, along with the order volumes from buyers and sellers at various price levels. Compared to historical stock line and candlestick charts, footprint charts offer richer detail, enabling more accurate assessments of breakout authenticity.

The output should include both the rationale and the answer as illustrated in Figure 1. This design is chosen because the investment field demands high explainability of decisions, and auditing the rationality behind decisions helps mitigate the risk of overvalued accuracy caused by guesses.

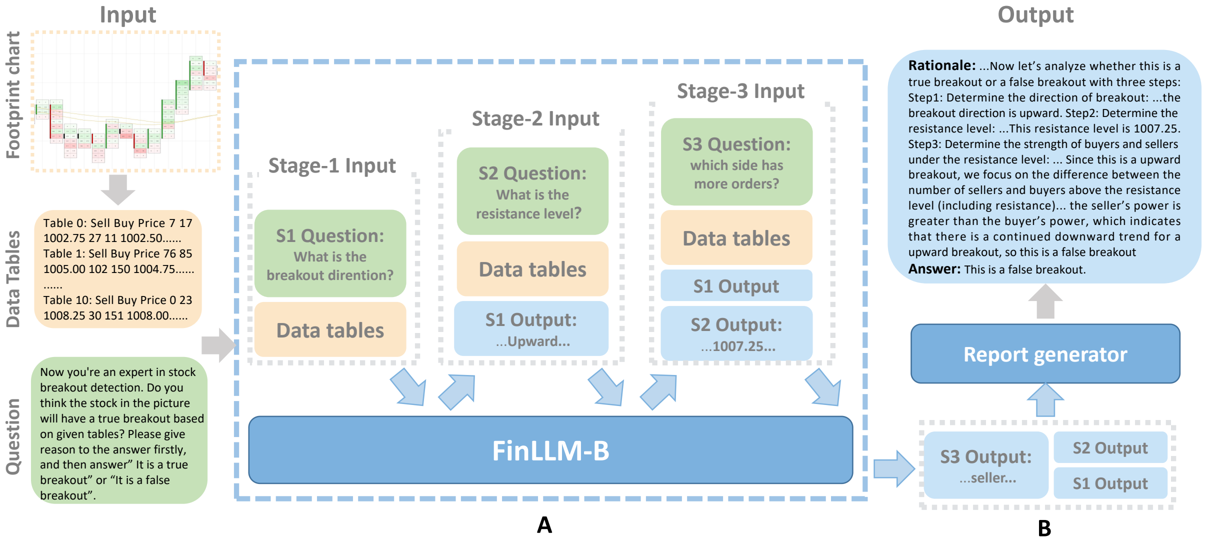


Table 0: Sell Buy Price 7 17 1002.75 27 11 1002.50......
Table 1: Sell Buy Price 76 85 1005.00 102 150 1004.75......
......
Table 10: Sell Buy Price 0 23 1008.25 30 151 1008.00......

Now you're an expert in stock breakout detection. Do you think the stock in the picture will have a true breakout based on given tables? Please give reason to the answer firstly, and then answer" It is a true breakout" or "It is a false breakout".

Rationale: ...Now let's analyze whether this is a true breakout or a false breakout with three steps: Step1: Determine the direction of breakout: ...the breakout direction is upward. Step2: Determine the resistance level: ...This resistance level is 1007.25. Step3: Determine the strength of buyers and sellers under the resistance level: ... Since this is a upward breakout, we focus on the difference between the number of sellers and buyers above the resistance level (including resistance)... the seller's power is greater than the buyer's power, which indicates that there is a continued downward trend for a upward breakout, so this is a false breakout
Answer: This is a false breakout.

Figure 2: Overview of FinLLM-B with multi-stage structure. Multi-stage structure consists of two parts: Part A and Part B. Part A comprises three stages, each corresponding to a subtask of breakout detection. Part B is responsible for integrating the answers from Part A into a rationale and providing the final answer.

4 Method

Our model is designed for financial breakout detection, with inputs being prompts and specialized data tables. The multi-stage architecture is the main framework of our model, as shown in Figure 2. The reasons for its design are as follows. The amount of data is limited in our task. Researchers usually choose to tackle this task by fine-tuning models directly. In the initial trial of our study, we attempted to address the problem by directly fine-tuning with one LLM as well, but the results were unsatisfactory. We think that reasoning and drawing conclusions are the two main steps humans take to solve this task. Based on this, we create two distinct datasets and trained two LLMs respectively responsible for reasoning and conclusion: FinLLM-B and report generator. Under this structure, FinLLM-B focuses on the problem itself rather than the details of report generation.

However, simply splitting the whole model into two parts for FinLLM-B and the report generator still has limited improvement. We find that longer outputs tend to increase errors. Therefore, based on the steps to solve the problem, we divide the training set for FinLLM-B into three parts, each part responsible for answering one subtask with a standard answer. This design offers three advantages. Firstly, this structure provides a framework for breakout detection, serving as prior knowledge to compensate for the lack of data. Secondly, these sub-tasks have a sequential relationship. They share parameters and complement each other so that we can more effectively solve these subtasks with one large language model (FinLLM-B). Thirdly, each part answers only one question, allowing it to focus on specialized knowledge and provide concise responses. This approach is similar to the division of labor and cooperation within a human team, significantly enhancing the accuracy and stability of final outputs.

4.1 Multi-Stage Structure

The model consists of two parts: task flow (Part A) and report generator (Part B), as shown in Figure 2.

Task Flow. The task flow primarily consists of three parts: Stage 1 (S1) task, Stage 2 (S2) task, and Stage 3 (S3) task, which correspond to the three steps of breakout detection as follows. Firstly, we need to determine the direction of the entire breakout. If the historical price shows an upward trend, it indicates an upward breakout. Secondly, the resistance level of the breakout needs to be identified. Identifying the resistance level depends on the direction of the breakout. For an upward breakthrough, its resistance level is the historical price's highest value, defined as the highest price point in the ten time units preceding the current time. For a downward breakout, its resistance level is the historical price's lowest value. Thirdly, we need to compare the forces of buyers and sellers, with the comparison point varying based on the results of the previous two steps. For an upward

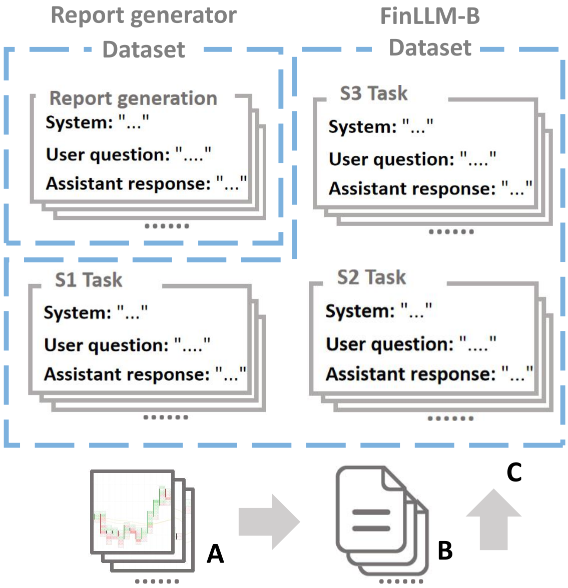


Figure 3: Dataset Construction. A: Footprint chart. B: Data table derived from the footprint chart. C: Dataset. It consists of two parts: FinLLM-B dataset and report generator dataset.

breakout, we compare the number of buy and sell orders above the resistance level, and vice versa for a downward breakout. The side with more orders is considered the stronger force.

FinLLM-B is employed to complete these three stages, providing evidence for breakout authenticity. It is pre-trained on GPT-3.5 and fine-tuned with 10 epochs for optimal performance.

Report Generator. The Report Generator is another large language model in our study. Its function is to aggregate the answers from FinLLM-B in sub-tasks and output an analysis report with the conclusion on the authenticity of the breakout. It fundamentally differs from FinLLM-B in functionality, hence it is trained independently on GPT-3.5, focusing exclusively on report generation.

4.2 Dataset

The process of dataset construction is shown in the Figure 3. The source data is collected as minute-level S&P 500 future footprint data from the NinjaTrader platform. We convert the source data into a special data table and then build the dataset. Compared to getting raw data directly from the platform, this approach saves 90% of the capital cost and provide better adaptability for LLMs. After obtaining the data tables, we use manual annotation to construct the dataset.The accuracy of human data annotation is ensured based on the expertise of the annotator and real market simulation.

FinLLM-B Dataset. This dataset involves two parts: training and testing. For training, The dataset consists of 60 training data. This includes 20 source data for each of S1, S2, and S3, and 10 samples for each of the true and false breakouts. For testing, the model will be tested a total of 1200 rounds, including 40 source data for each stage, and each test is repeated 10 times to test the stability of the model. Each training data consists of three parts based on the official setup of OpenAI: system, user question, and assistant response.

Report Generator Dataset. The dataset of report generator is simpler because its task is not complex. It has 20 training data, which are annotated by experts according to the task, and other settings are consistent with FinLLM-B.

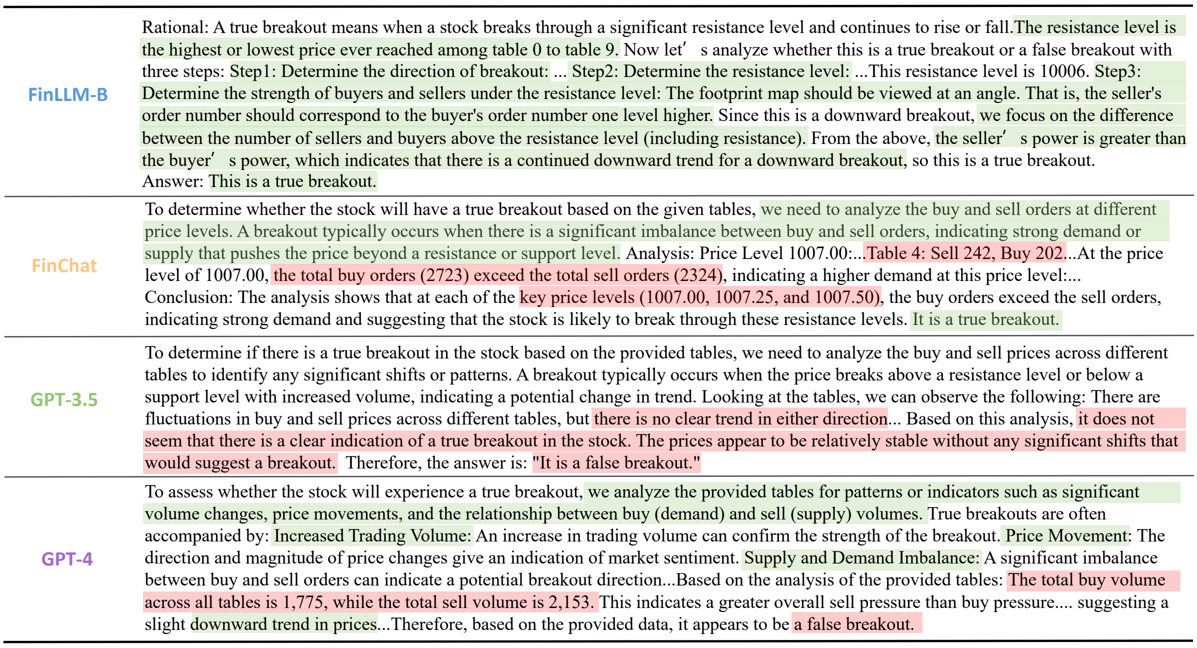
5 Experiment

5.1 Baseline & Evaluation Metrics.

FinLLM-B was trained based on GPT-3.5 and compared with four baselines: GPT-3.5 (OpenAI, 2022), GPT-4 (Achiam et al., 2023), FinChat (FinChat, 2024), and Long Short-Term Memory network (LSTM) (Bhandari et al., 2022). FinChat is a commercial-grade financial LLM that adapts GPT specifically for the finance sector. LSTM is a special recurrent neural network which is frequently used for financial prediction. We evaluated FinLLM-B in three main aspects: professionalism, accuracy, and stability.

Professionalism evaluation. Since evaluating the expertise of the model's responses is subjective, we used manual scoring by professionals to assess the expertise of models.

Accuracy evaluation. We compared the accuracy rate and perfection rate of each model. The accuracy rate is derived from the statistical analysis of the model's actual results. In addition, to evaluate the performance of the final report, we introduced the perfection rate, representing the proportion of samples that produced entirely accurate reports out of all test samples. An entirely accurate report correctly identifies each sub-task, the reasoning process, and the final breakout judgment. The calculation method is: S1 accuracy * S2 accuracy * S3 accuracy. Under this evaluation criterion, if the real market result is a true breakout, but the tested model's answer is a true breakout with incorrect



FinLLM-B	Rational: A true breakout means when a stock breaks through a significant resistance level and continues to rise or fall.The resistance level is the highest or lowest price ever reached among table 0 to table 9. Now let′ s analyze whether this is a true breakout or a false breakout with three steps: Step1: Determine the direction of breakout: ... Step2: Determine the resistance level: ...This resistance level is 10006. Step3: Determine the strength of buyers and sellers under the resistance level: The footprint map should be viewed at an angle. That is, the seller's order number should correspond to the buyer's order number one level higher. Since this is a downward breakout, we focus on the difference between the number of sellers and buyers above the resistance level (including resistance). From the above, the seller′ s power is greater than the buyer′ s power, which indicates that there is a continued downward trend for a downward breakout, so this is a true breakout. Answer: This is a true breakout.
FinChat	To determine whether the stock will have a true breakout based on the given tables, we need to analyze the buy and sell orders at different price levels. A breakout typically occurs when there is a significant imbalance between buy and sell orders, indicating strong demand or supply that pushes the price beyond a resistance or support level. Analysis: Price Level 1007.00:...Table 4: Sell 242, Buy 202...At the price level of 1007.00, the total buy orders (2723) exceed the total sell orders (2324), indicating a higher demand at this price level:... Conclusion: The analysis shows that at each of the key price levels (1007.00, 1007.25, and 1007.50), the buy orders exceed the sell orders, indicating strong demand and suggesting that the stock is likely to break through these resistance levels. It is a true breakout.
GPT-3.5	To determine if there is a true breakout in the stock based on the provided tables, we need to analyze the buy and sell prices across different tables to identify any significant shifts or patterns. A breakout typically occurs when the price breaks above a resistance level or below a support level with increased volume, indicating a potential change in trend. Looking at the tables, we can observe the following: There are fluctuations in buy and sell prices across different tables, but there is no clear trend in either direction... Based on this analysis, it does not seem that there is a clear indication of a true breakout in the stock. The prices appear to be relatively stable without any significant shifts that would suggest a breakout. Therefore, the answer is: "It is a false breakout."
GPT-4	To assess whether the stock will experience a true breakout, we analyze the provided tables for patterns or indicators such as significant volume changes, price movements, and the relationship between buy (demand) and sell (supply) volumes. True breakouts are often accompanied by: Increased Trading Volume: An increase in trading volume can confirm the strength of the breakout. Price Movement: The direction and magnitude of price changes give an indication of market sentiment. Supply and Demand Imbalance: A significant imbalance between buy and sell orders can indicate a potential breakout direction...Based on the analysis of the provided tables: The total buy volume across all tables is 1,775, while the total sell volume is 2,153. This indicates a greater overall sell pressure than buy pressure.... suggesting a slight downward trend in prices...Therefore, based on the provided data, it appears to be a false breakout.

Figure 4: Output samples for professionalism comparison. Green: Valuable domain knowledge. Red: Incorrect domain knowledge and mistakes. Due to the length of the output, we used '...' to omit non-essential content.

Models	S1 Accuracy	S2 Accuracy	S3 Accuracy	Average Accuracy	Perfection Rate
GPT-3.5	50.25 ± 10.30	10.50 ± 5.99	41.50 ± 10.55	34.83	2.19
GPT-4	61.50 ± 8.83	13.50 ± 4.74	52.25 ± 6.71	42.42	4.34
FinChat	75.5 ± 8.96	23.25 ± 9.86	60.50 ± 5.99	53.42	11.18
LSTM	–	–	–	–	45
FinLLM-B (Ours)	$\mathbf{95.00 \pm 0.00}$	$\mathbf{89.40 \pm 8.72}$	$\mathbf{70.00 \pm 0.00}$	**84.80**	**59.45**

Table 1: Result highlights. Accuracy and perfection rates of FinLLM-B and baseline models are evaluated based on correct identification of sub-tasks, reasoning process, and final breakout judgment. Note: LSTM only provides final results which are considered as the perfection rate.

reasoning, we consider the report inaccurate. This calculation method is necessary because having only the answers does not adequately reflect the model's capability.

Stability Evaluation. Two testing methods were used to evaluate the stability of the model: standard deviation and output consistency distribution. For standard deviation, each model was tested 1200 rounds in total. We tested 40 sets of samples for task S1-3, each repeated 10 times and recorded each result for calculating the standard deviation. For output consistency distribution, we tested 40 sets of samples, each set tested 10 times repeatedly, and recorded the quantity of samples which produced same outputs across repeated tests. Specifically, we recorded the number of samples with 100% same, 80% same, 60% same, and less than 60% same. For example, if a test sample produces consistent outputs 8 times out of 10 repeated outputs, it is recorded as 80% same in this round of testing. We are particularly concerned with cases where the outputs are 100% same, indicating that the sample produced the same output all 10 times, demonstrating high reliability. We used the output consistency distribution because results of breakout detection will be used for investment decisions, thus requiring high consistency.

5.2 Main Results

FinLLM-B outperforms other LLMs and neural network models, as shown in Table 1. It surpasses GPT-3.5 by 49.97% in average accuracy and 57.26% in perfection rate, primarily due to the baseline models' lower performance in the S2 task.

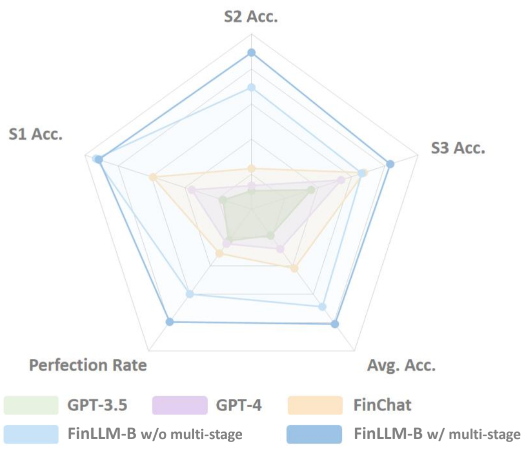


Figure 5: Accuracy comparison. Each axis is rescaled independently for better comparison.

FinLLM-B	w/o multi-stage	w/ multi-stage
S1 Accuracy	96.00 ± 1.75	95.00 ± 0.00
S2 Accuracy	69.50 ± 5.63	89.40 ± 8.72
S3 Accuracy	59.75 ± 5.47	70.00 ± 0.00
Average Accuracy	75.08	**84.80**
Perfection Rate	39.87	**59.45**

Table 2: Accuracy comparison between FinLLM-B with and without multi-stage. The proposed multi-stage structure demonstrates a notable improvement in the accuracy and perfection rate.

5.3 Report Generator

We assessed the report generator's performance using 40 test samples, each tested 10 times. The generator consistently achieved expected results, due to the relatively simple nature of the task.

Professionalism. Scoring results reveal that FinLLM-B scored the highest, with an average of 8 out of 10, compared to GPT-4 and FinChat (6 out of 10) and GPT-3.5 (3 out of 10). Test samples shown in Figure 4 indicate that FinLLM-B demonstrates a clearer structure, more stable performance, and superior reasoning capabilities than the baselines.

Accuracy. Figure 5 and Table 2 illustrates that FinLLM-B achieves significantly higher accuracy than other LLMs, especially in task S2. S2 task better highlights the model's strengths due to its uncountable answer space, unlike the countable answers in S1 and S2, where guessing inflates accuracy. LSTM's accuracy, close to 50%, is limited by its requirement on substantial training data, which is difficult to obtain in our task.

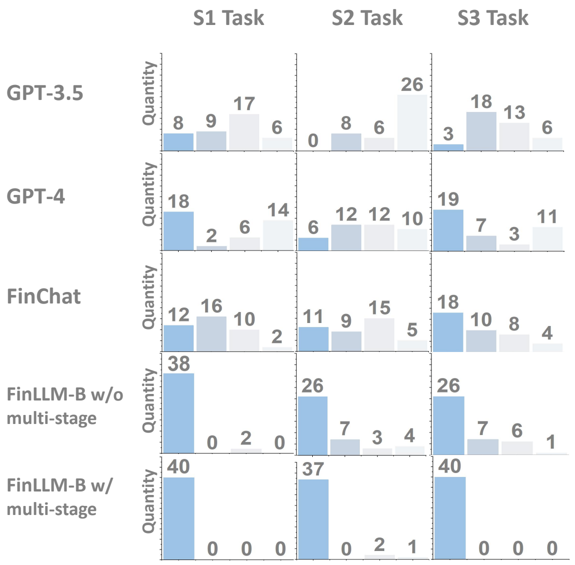


Figure 6: Output consistency distribution. Blue areas represent better stability. It represents the quantity of samples that have all same outputs in the stability test.

Models	S1	S2	S3
GPT-3.5	0.37	170.05	0.43
GPT-4	0.25	0.39	0.23
Finchart	0.29	0.32	0.23
FinLLM-B w/o multi-stage	0.02	0.14	0.16
FinLLM-B w/ multi-stage	**0.00**	**0.06**	**0.00**

Table 3: Standard deviation. Actual resistance level values are used to calculate the standard deviation in S2.

Stability. Figure 6 and Table 3 highlight FinLLM-B's stability advantages, particularly in S2. GPT-3.5's performance in S2 is significantly low. This is because the standard deviation here is the actual result's standard deviation, and GPT-3.5 often outputs values significantly different from the actual result. In Figure 6, the blue area indicates the number of samples with all same output in 10 tests, demonstrating the stability of FinLLM-B. GPT-3.5 frequently switches between two answers, indicating that its accuracy is based on guessing.

5.4 Report Generator

We assessed the report generator's performance using 40 test samples, each tested 10 times. The generator consistently achieved expected results, due to the relatively simple nature of the task.

5.5 Ablation Study

We compared FinLLM-B with and without the multi-stage structure, as shown in Figures 5-6 and Tables 2-3. Two key findings emerged: 1) The multi-stage structure significantly improves accuracy, particularly in S2. 2) Stability is enhanced with the multi-stage design. These improvements arise from the structure's design. Under the multi-stage structure, the report generator handles report creation, allowing FinLLM-B to focus on answering questions. Each of three components in FinLLM-B specializes in a specific aspect, sharing parameters to enhance accuracy and stability.

5.6 Dataset Size Analysis

We tested the model's accuracy with different dataset sizes and found that the current 10 shots scale is appropriate. Samples in the dataset are categorized into two types: true and false breakout. We expanded the training set by increasing both the true and false breakout samples. For every 2 shots increase, we recorded the model's accuracy based on a single test run. The model accuracy for 2 to 10 shots is as follows: 57.50%, 70.83%, 78.21%, 82.87%, and 84.80%. From the records, the rate of accuracy improvement slows down, and the rising trend curve becomes nearly flat at 10 shots. This indicates that our model's performance can improve with more training data, and at 10 shots, the performance is nearing its peak, suggesting that a 10-shot size is appropriate. Additionally, the model performed well with only 10 shots, further indicating that using LLMs is a promising approach for breakout detection in data-limited scenarios.

6 Future Work

Our work is the first to explore the application of large language models in financial breakout detection tasks, and we propose a multi-stage framework that enables our model to outperform other competitors. However, there is still room for improvement in the following two aspects.

Future work could expanding data modalities, such as images or videos, to better align the model with real-world scenarios. Currently, FinLLM-B relies on minute-level data from converted static footprint charts. However, the financial trading market changes rapidly, and continuous dynamic data could improve breakout detection accuracy. For instance, FinLLM-B could directly input videos to capture real-time changes in buy and sell orders in the future, enhancing breakout detection performance. Additionally, enriching the dataset with a broader range would provide deeper insights into the model's optimal performance and robustness.

There is still room for improvement in the accuracy of the S3 task. We found the accuracy of S3 is significantly lower than the other two subtasks primarily due to its inherent complexity. The S3 task involves comparing the strength of buyers and sellers based on resistance levels, a process that is relatively intricate. This complexity may limit the full utilization of the capabilities of large language models. In the future, researchers could further segment the S3 task using a multi-stage structure to attempt to improve its accuracy.

7 Conclusion

We present FinLLM-B, the first large language model specifically designed for breakout detection, which alleviates the important issue in financial breakout trading field. To develop this model, we construct a high-quality financial breakout dataset. Furthermore, we create an innovative multi-stage framework, distinguishing FinLLM-B from the report generator and segmenting it into three distinct components based on problem-solving steps. This design enables FinLLM-B to more effectively demonstrate domain knowledge and enhances the model's accuracy and stability in our task. We believe that our model will serve as a valuable resource for future research and foster further exploration in the field of financial breakout trading.

Acknowledgement

Weiran Huang is funded by the National Natural Science Foundation of China (No. 62406192), Opening Project of the State Key Laboratory of General Artificial Intelligence (No. SKLAGI2024OP12), Tencent WeChat Rhino-Bird Focused Research Program, and Doubao LLM Fund.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ray Ball. 2009. The global financial crisis and the efficient market hypothesis: what have we learned? *Journal of Applied Corporate Finance*, 21(4):8–16.

Daniel Ben David, Yehezkel S Resheff, and Talia Tron. 2021. Explainable ai and adoption of financial algorithmic advisors: an experimental study. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 390–400.

Hendrik Bessembinder and Kalok Chan. 1995. The profitability of technical trading rules in the asian stock markets. *Pacific-basin finance journal*, 3(2-3):257–284.

Hum Nath Bhandari, Binod Rimal, Nawa Raj Pokhrel, Ramchandra Rimal, Keshab R Dahal, and Rajendra KC Khatri. 2022. Predicting stock market index using lstm. *Machine Learning with Applications*, 9:100320.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2023. Oceangpt: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*.

Lawrence Blume, David Easley, and Maureen O'hara. 1994. Market statistics and technical analysis: The role of volume. *The journal of finance*, 49(1):153–181.

Al Brooks. 2011. *Trading Price Action Trading Ranges: Technical Analysis of Price Charts Bar by Bar for the Serious Trader*. John Wiley & Sons.

Tom 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*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. 2002. Order imbalance, liquidity, and market returns. *Journal of Financial economics*, 65(1):111–130.

Alexander Elder. 2002. *Come into my trading room: A complete guide to trading*, volume 146. John Wiley & Sons.

FinChat. 2024. Finchat: Ai-powered financial chatbot. https://finchat.io/. Accessed: 2024-11-27.

Sebastian Fritz-Morgenthal, Bernhard Hein, and Jochen Papenbrock. 2022. Financial risk management and explainable, trustworthy, responsible ai. *Frontiers in artificial intelligence*, 5:779799.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Thomas F Gosnell, Arthur J Keown, and John M Pinkerton. 1996. The intraday speed of stock price adjustment to major dividend changes: Bid- ask bounce and order flow imbalances. *Journal of Banking & Finance*, 20(2):247–266.

Chenyu Han and Xiaoyu Fu. 2023. Challenge and opportunity: deep learning-based stock price prediction by using bi-directional lstm model. *Frontiers in Business, Economics and Management*, 8(2):51–54.

Ancy John and T Latha. 2023. Stock market prediction based on deep hybrid rnn model and sentiment analysis. *Automatika*, 64(4):981–995.

Taewook Kim and Ha Young Kim. 2019. Forecasting stock prices with a feature fusion lstm-cnn model using different representations of the same data. *PloS one*, 14(2):e0212320.

Timothy Knight. 2010. *Chart Your Way to Profits: The Online Trader's Guide to Technical Analysis with ProphetCharts*, volume 475. John Wiley & Sons.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2024. Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32.

Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11(10):2320.

Camillo Lento, Nikola Gradojevic, et al. 2007. The profitability of technical trading rules: A combined signal approach. *Journal of Applied Business Research (JABR)*, 23(1).

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llavamed: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Yuesen Li, Chengyi Gao, Xin Song, Xiangyu Wang, Yungang Xu, and Suxia Han. 2023. Druggpt: A gpt-based strategy for designing potential ligands targeting specific proteins. *bioRxiv*, pages 2023–06.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519.

Andrew W Lo, Harry Mamaysky, and Jiang Wang. 2000. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The journal of finance*, 55(4):1705–1765.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Thorben Lubnau and Neda Todorova. 2014. Technical trading revisited: evidence from the asian stock markets. *Corporate Ownership & Control*, 11(2):511–532.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Burton G Malkiel. 2003. The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1):59–82.

Timothy R McIntosh, Tong Liu, Teo Susnjak, Paul Watters, Alex Ng, and Malka N Halgamuge. 2023. A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence*.

John J Murphy. 1999. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin.

OpenAI. 2022. Chatgpt. `https://openai.com/blog/chatgpt/`. Accessed: 2024-06-14.

Mahendra Raj and David Thurston. 1996. Effectiveness of simple technical trading rules in the hong kong futures markets. *Applied Economics Letters*, 3(1):33–36.

Eric Sarrion. 2023. The implications of chatgpt on employment and society. In *Exploring the Power of ChatGPT: Applications, Techniques, and Implications*, pages 73–82. Springer.

Lynn A Stout. 2002. The mechanisms of market inefficiency: An introduction to the new finance. *J. Corp. L.*, 28:635.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*.

Mark P Taylor and Helen Allen. 1992. The use of technical analysis in the foreign exchange market. *Journal of international Money and Finance*, 11(3):304–314.

Michael Valtos. 2015. *Trading Order Flow*. Orderflows. Retrieved from `https://www.orderflows.com/book/TradingOrderFlow768.pdf`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Jilin Zhang, Lishi Ye, and Yongzeng Lai. 2023a. Stock price prediction using cnn-bilstm-attention model. *Mathematics*, 11(9):1985.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *Preprint*, arXiv:2210.03493.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023c. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zihao Zhang, Stefan Zohren, and Roberts Stephen. 2020. Deep reinforcement learning for trading. *The Journal of Financial Data Science*.

Hong Zhu, Zhi-Qiang Jiang, Sai-Ping Li, and Wei-Xing Zhou. 2015. Profitability of simple technical trading rules of chinese stock exchange indexes. *Physica A: Statistical Mechanics and its Applications*, 439:75–84.