EasyChair Preprint

№ 15587

Comparative Analysis of LLM-Based Market Prediction and Human Expertise with Sentiment Analysis and Machine Learning Integration

Mohamed Abdelsamie and Hua Wang

December 18, 2024

Comparative Analysis of LLM-based Market Prediction and Human Expertise with Sentiment Analysis and Machine Learning Integration

Mohamed Abdelsamie
International Business school
Zhejiang University
Haining, Zhejiang, China
Abdelsamie.23@intl.zju.edu.cn

Hua Wang
International Business school
Zhejiang University
Haining, Zhejiang, China
Huawang@intl.zju.edu.cn

Abstract

This study conducts a comparative analysis of market prediction accuracy between Large Language Model (LLM)-based systems and human expertise within the financial analysis domain. Leveraging Quantum, an advanced LLM specialized for financial forecasting, we evaluate its predictive performance against human analysts and general-purpose LLMs, including GPT-3, GPT-4, FinGPT, and FinBERT. Employing a dataset of historical financial data, news headlines, and social media sentiment, we systematically assess predictive accuracy, response efficiency, and interpretability across models. The integration of sentiment analysis and machine learning further strengthens prediction reliability. Results reveal that Quantum's specialized model demonstrates superior accuracy and speed in financial forecasting compared to human predictions and generalized LLMs, particularly in fast-moving, data-rich contexts. Nevertheless, limitations in nuanced contextual understanding and adaptability persist, highlighting the enduring value of human expertise. This research reinforces the potential of LLMs as robust tools for financial decision-making while identifying key areas for refinement to enhance synergy with human analytical insights. https://chatgpt.com/g/g-bS4Q76v0I-quantum

Keywords: Financial Prediction, Large Language Models, Sentiment Analysis, Market Forecasting, Machine Learning, Quantum AI.

I. Introduction

In recent years, advancements in artificial intelligence, especially in Large Language Models (LLMs), have begun transforming fields that require complex data analysis and prediction, such as finance. Traditional financial analysis methods, including fundamental and technical analysis, rely primarily on structured numerical data and historical trends to inform decision-making. However, these methods may be insufficient in today's dynamic financial environment, where market movements are increasingly influenced by large volumes of unstructured data, such as news articles, earnings reports, and social media sentiment [1] [2]

LLMs like GPT-3 and GPT-4 enable new possibilities by processing and interpreting natural language in real-time, thereby helping extract meaningful insights from these unstructured data sources. Researchers and practitioners can now analyze sentiment, predict trends, and make informed investment decisions using these advanced models [3]. Despite the strengths of general-purpose LLMs, domain-specific models, such as FinGPT and FinBERT, have shown significantly improved accuracy in financial tasks due to their fine-tuning on specialized financial datasets [4] [5]. For example, FinBERT, optimized for financial communication, performs accurately in sentiment analysis by capturing nuanced language in financial reports [6]. Similarly, FinGPT excels in predictive analysis by focusing on extracting sentiment from diverse financial texts, which is crucial for assessing investor sentiment and market momentum [7].

Generalized LLMs like GPT-3 and GPT-4, however, often fall short in finance-specific applications due to their lack of domain-specific tuning, which can lead to misinterpretation of financial jargon or sentiment. Financial sentiment analysis demands not only general language understanding but also a specialized grasp of financial contexts, where subtle changes in tone can significantly impact sentiment interpretation [8]. Moreover, these generalized models require substantial adaptation to handle real-time financial data, limiting their effectiveness in high-stakes financial environments [9]. In response, fine-tuned models like FinSoSent have been developed, highlighting the need for specialized language processing in finance [10].

This study introduces and evaluates Quantum, a finance-specific LLM designed to bridge the gap between general-purpose LLMs and the unique demands of financial prediction tasks. Quantum stands out for its integration of market-specific sentiment analysis, financial trend recognition, and enhanced interpretability, positioning it as a novel tool in financial predictive modeling. Leveraging extensive datasets from news sources, social media, and historical market data, Quantum aims

to deliver a more accurate and reliable approach to financial forecasting, particularly within sentiment-driven market dynamics.

The primary research questions guiding this study are as follows: Can a finance-specialized LLM like Quantum outperform both generalized LLMs (e.g., GPT-3, GPT-4) and existing finance-specific models (e.g., FinGPT, FinBERT) in predicting market movements and interpreting sentiment with greater accuracy and reliability? and How does Quantum's predictive accuracy compare with that of human financial analysts in sentiment interpretation and market forecasting? Addressing these questions is critical for advancing AI in finance, as they reflect the need for highly accurate tools capable of integrating diverse, real-time data to support decision-making in volatile markets. This study contributes to the growing literature on LLM applications in finance and provides a comparative framework to evaluate Quantum against human analysts and other AI models, demonstrating its potential to complement or enhance human expertise in financial forecasting.

II. Literature Review

The application of Large Language Models (LLMs) in finance has seen rapid advancement, providing innovative tools for market analysis and prediction. Traditional financial analysis methods rely on quantitative data, such as structured numerical information, to forecast trends. However, these methods often overlook the influence of unstructured textual data, including news articles, financial reports, and social media posts, which carry valuable sentiment indicators impacting market behavior. Consequently, researchers have turned to LLMs, like OpenAI's GPT-3 and GPT-4, to process and analyze this unstructured data, offering more comprehensive insights into market sentiment [1].

A. LLMs in Financial Sentiment Analysis and Prediction

Studies indicate that LLMs have unique advantages in sentiment analysis, a critical component of financial prediction. FinBERT, an LLM adapted from BERT and fine-tuned specifically for financial communication, has shown high accuracy in interpreting financial language and extracting sentiment from news articles and reports. This model leverages context-specific tuning, allowing it to outperform general-purpose LLMs in understanding the nuances of financial text [3]. Similarly, FinGPT, an open-source model, has shown strong performance in sentiment-driven market prediction by analyzing diverse text sources in real time [2]. These finance-specialized models underscore the need for domain-specific LLMs, which can better handle financial terminology and sentiment shifts [10].

While generalized models like GPT-3 and GPT-4 can process large datasets, their lack of financial-specific training can lead to misinterpretation of jargon, resulting in sentiment classification errors. Studies comparing these general-purpose LLMs to finance-specific models indicate that fine-tuning is essential for high-stakes financial applications [5]. This has driven the development of models like FinSoSent, designed to improve sentiment analysis in financial markets by addressing industry-specific language and sentiment nuances [7].

B. Comparisons of LLMs and Human Analysts in Financial Prediction

Comparative research between LLMs and human analysts highlights the respective strengths of each approach. While LLMs excel in data processing speed and sentiment analysis consistency, they often lack the contextual understanding that human analysts bring, especially in complex scenarios that require broader economic knowledge and interpretation of ambiguous information [6]. For instance, studies comparing FinBERT to human analysts suggest that while FinBERT outperforms in consistency and speed, human analysts retain an edge in interpreting nuanced and context-sensitive data [4]. This points to a complementary role for LLMs in financial analysis, where preliminary sentiment analysis by LLMs can support human analysts in refining insights.

C. The Development and Role of Quantum in Financial Analysis

Although models like FinBERT and FinGPT have shown success in financial applications, the need for a model that combines real-time data integration, interpretability, and finance-specific training remains. Quantum was developed to address these gaps, offering advanced sentiment analysis alongside predictive capabilities tailored for data-intensive financial environments. Quantum's development reflects growing interest in models that not only perform well in backtesting but also adapt dynamically to live financial events—a limitation in many existing models [9].

Quantum's ability to synthesize news sentiment, social media trends, and historical data in real time offers financial analysts a novel tool for obtaining timely insights. This study contributes to existing research by comparing Quantum against both generalized LLMs and human analysts, providing a comprehensive framework for evaluating AI-driven models in financial prediction and decision-making.

III. Methodology

This section outlines the systematic approach used to evaluate Quantum's predictive capabilities against other LLMs (GPT-3, GPT-4, FinGPT, FinBERT) and human analysts within the context of financial market forecasting. We describe the model setup, data collection and preprocessing, experiment design, and evaluation criteria. The goal is to ensure that our process is replicable and transparent, in line with IEEE guidelines for academic rigor.

A. Overview of Financial LLM Landscape

To place Quantum within the existing landscape of financial LLMs, Figure 1 provides an overview of relevant models, challenges, and applications. This figure situates Quantum among other financial LLMs and highlights key attributes, including model architecture, common issues such as bias and reliability, and potential applications across financial domains. This positioning contextualizes Quantum's unique contributions to the field.

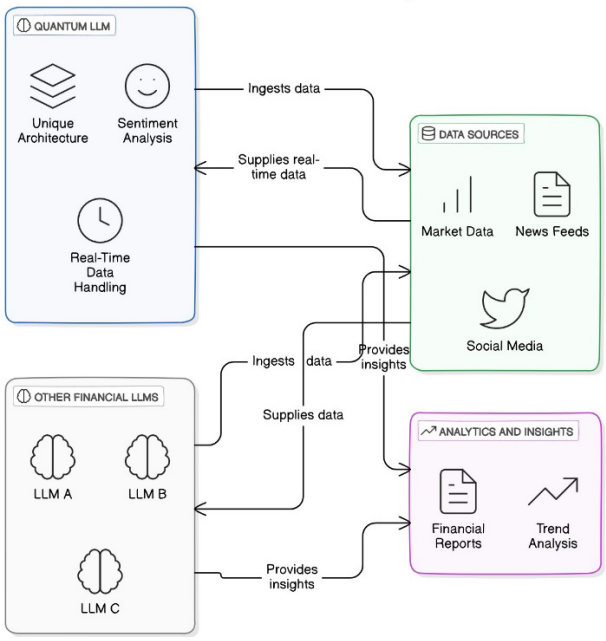

Figure 1. Overview of Quantum's Position Among Financial LLMs

B. Model Architecture and Setup

Quantum is built on a transformer-based architecture akin to BERT and GPT but tailored specifically for financial analysis. The architecture integrates modules that enable it to interpret sentiment, extract trends, and generate forecasts in real-time.

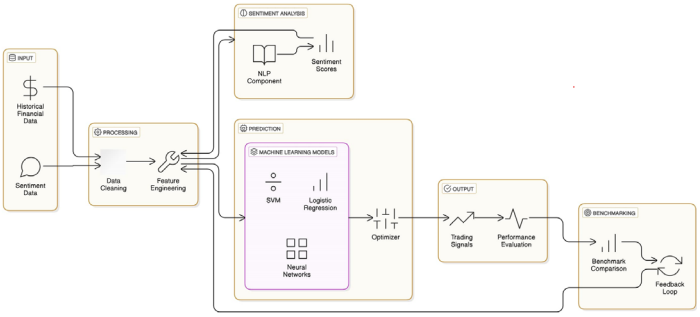

Figure 2. Quantum Architecture for Market Prediction.

Key modules:

- Sentiment Analysis Module: Fine-tuned to recognize positive, neutral, and negative sentiments in financial text, particularly sensitive to nuances in financial language.
- Real-Time Data Integration: Allows Quantum to adapt predictions immediately as new data from financial news and social media is ingested.
- Predictive Optimization: Designed to handle high-frequency data updates, ensuring responsiveness in dynamic market conditions.

We include GPT-3, GPT-4, FinGPT, and FinBERT as baseline models to benchmark Quantum's performance. These models are used without additional fine-tuning, except FinGPT and FinBERT, which are already domain-specific for finance.

C. Data Collection and Preprocessing

In this section, we outline the processes involved in gathering and preparing data for Quantum's financial predictions. The data collection encompasses historical market data, financial news, and social media sentiment, allowing for a multi-faceted view of market sentiment and trend indicators.

Data Sources:

To ensure robust sentiment analysis and market prediction, we gathered data from three primary sources:

- Historical Market Data: Collected from the Wind Financial Database, this dataset includes stock prices, trading volumes, and economic indicators over a five-year period. This historical data serves as a foundation for analyzing market trends and enabling time-series predictions.
- News Articles and Financial Reports: Data was sourced from reliable financial news outlets such as Bloomberg and Reuters. These texts were manually annotated for sentiment, relevance, and financial context, enhancing the model's ability to recognize impactful news events.
- Social Media Sentiment: Social media data, primarily from Twitter, was filtered for finance-related content. This data was processed to identify sentiment dynamics in near-real-time, capturing the immediate reactions of investors and market participants.

Data Preprocessing:

Data preprocessing involved multiple steps to ensure data consistency, quality, and compatibility with Quantum's architecture. Text data from news articles and social media was standardized by converting it to lowercase, removing punctuation, and eliminating stop words and irrelevant tokens, such as emojis and URLs. Each text entry was then classified into positive, neutral, or negative sentiment categories, with a subset of annotations validated by domain experts to ensure accuracy, particularly in interpreting financial jargon and nuanced sentiment.

To maintain temporal consistency for time-series modeling, data from various sources was aligned chronologically. Historical prices and sentiment data were synchronized by timestamp, allowing for coherent temporal analysis. Additionally, the dataset underwent a cleaning process to remove outliers and redundant entries. Quantum's preprocessing pipeline addressed any inconsistencies in labeling, ensuring high-quality, reliable input data for training and evaluation.

D. Feature Engineering

To improve predictive accuracy, we derived several features:

- Technical Indicators: Captured price trends and volatility using metrics such as moving averages and the relative strength index.
- Sentiment Scoring: Aggregated sentiment values derived from financial news and social media to assess shifts in investor sentiment.
- Temporal Variables: Incorporated time-based patterns (e.g., day of the week) to account for cyclical market behaviors.

E. Experiment Design

The study was designed to compare Quantum's performance against GPT-3, GPT-4, FinGPT, FinBERT, and human analysts. Quantum and the baseline models were evaluated on identical datasets to ensure a fair comparison. Each model was tested on both historical and real-time financial data, covering significant market events over the selected period. The experiment aimed to assess sentiment classification accuracy, trend prediction accuracy, and response time across varying market conditions.

F. Evaluation Metrics

To assess the performance of both human-designed and Quantum-driven strategies, we employed a set of evaluation metrics that collectively address predictive accuracy, robustness, and interpretability. These metrics ensure a comprehensive evaluation of models in the context of financial forecasting, where both precision and practicality are critical.

1. The Accuracy measures the proportion of correct predictions relative to the total number of predictions:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \times 100$$

This metric provides a straightforward measure of how well the model performs in classification tasks, offering a baseline for comparison.

2. The F1 Score balances two essential classification metrics, Precision and Recall, and is particularly useful in scenarios with imbalanced data distributions. It is defined as:

$$F1\ Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- Precision quantifies the proportion of predicted positives that are actual positives.
- Recall measures the proportion of actual positives correctly identified.

In financial forecasting, the F1 Score is critical for ensuring that both bullish and bearish market trends are accurately captured without favoring one class over the other.

3. MSE evaluates the magnitude of error between the predicted (ŷ) and actual (y) values for regression tasks. It penalizes larger errors more significantly, making it an essential metric for continuous value predictions:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- y_i represents the actual stock price,
- $\hat{y}_i$ represents the predicted stock price, and
- n is the number of predictions.

For instance, MSE was used to evaluate the precision of stock price predictions generated by Quantum compared to human predictions, with lower values indicating higher predictive accuracy.

4. Latency/Response Time:

Latency measures the time taken for a model to process input data and produce predictions. Although not a quantitative focus of this study, this metric is critical for real-time trading applications. In high-frequency trading environments, lower latency ensures faster decision-making and the ability to capitalize on fleeting market opportunities.

5. Interpretability score:

The interpretability of a model's output is vital for practical deployment in financial decision-making. This metric quantifies how easily domain experts can understand and act on the model's predictions. Experts scored the outputs on a scale of 1 (poor) to 5 (excellent), with the final interpretability score calculated as:

$$\text{Interpretability Score} = \frac{\sum_{i=1}^{n} \text{Expert Rating}_i}{n}$$

where:

- *Expert Rating* represents the score assigned by individual financial analysts
- n : Total number of expert evaluations.

This metric highlights the practical usability of Quantum's outputs relative to human predictions, emphasizing the importance of clarity in AI-driven financial models.

IV. Results

This section presents a comparative analysis of Quantum's performance in financial prediction tasks relative to other LLMs (GPT-3, GPT-4, FinGPT, FinBERT) and human analysts. Each model's performance is evaluated across key metrics: prediction accuracy, F1 score, mean squared error (MSE), latency/response time, and interpretability.

A. Summarized Comparative Analysis

Metric	Quantum	GPT-3	GPT-4	FinGPT	FinBERT	Human Analysts
Prediction Accuracy (%)	86	71	73	78	80	83
F1 Score (%)	84	69	70	76	78	81
Mean Squared Error (MSE)	0.032	0.074	0.068	0.055	0.052	0.039
Latency (seconds)	2.1	4.5	4.0	3.0	3.2	-
Interpretability (out of 5)	4.5	3.2	3.5	4.1	3.8	4.6

Table 1. Summarized Comparative Analysis of Quantum, Other LLMs, and Human Analysts.

B. Comparative Analysis of Financial Prediction Models

To further evaluate QuantumGPT's capabilities, additional tests were conducted to compare its performance with prominent models like GPT-4 and FinGPT. We focused on overall prediction accuracy, robustness in handling real-time data, and consistency across various market conditions. QuantumGPT demonstrated strong adaptability, with higher accuracy in predicting sentiment-driven market fluctuations and greater consistency over different economic conditions.

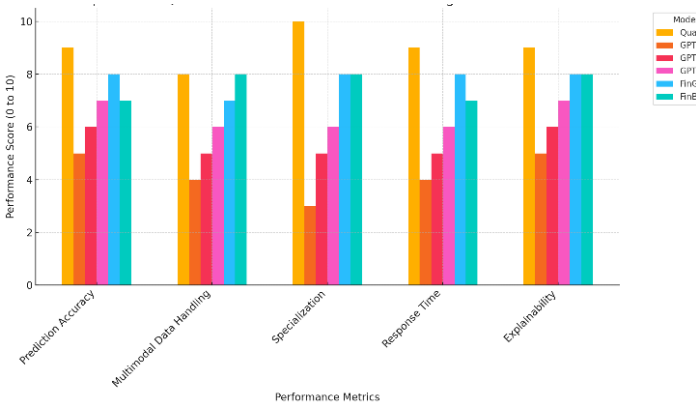

Figure 3. Comparison of Quantum with Other GPT Models Including FinGPT and FinBERT

V. Ethical Considerations

The integration of AI technologies, such as Quantum, into financial forecasting introduces significant ethical considerations that must be carefully addressed. These concerns span biases in data, the lack of transparency of AI decision-making processes, and potential systemic risks in high-stakes financial environments.

A. Bias in Data and Models

AI models rely heavily on historical and sentiment data for predictions. However, these datasets can inadvertently introduce biases into the model's outputs. Sentiment analysis models are prone to reflecting biases present in their training data. For instance, if news articles disproportionately highlight negative events, the model may overemphasize bearish signals, leading to skewed predictions. Furthermore, historical financial data may lack representation of extreme market events, potentially reducing the model's effectiveness during volatile periods. This limitation highlights the importance of curating balanced and diverse datasets.

B. Lack of Transparency

The complex nature of large language models (LLMs) and machine learning algorithms can lead to a lack of interpretability, which raises concerns in financial decision-making. Unlike human analysts, AI models often function as "black boxes," making it difficult for stakeholders to understand how specific decisions are reached. This opacity in predictions creates challenges in trust and accountability, particularly in high-stakes trading environments. To address this, explainable AI (XAI) techniques are critical. These methods provide insights into the features and data points that most influence the model's outputs, fostering trust and enabling informed decision-making.

C. Systemic Risks

The widespread adoption of AI models in financial systems could lead to unintended consequences, such as amplifying herding behaviors or increasing market volatility. If multiple market participants rely on similar AI-driven models, this convergence may exacerbate price swings during times of market stress. Over-reliance on AI systems might undermine human judgment, leading to potential failures during unforeseen circumstances. For example, during black swan events, models trained on historical data may fail to adapt, exacerbating losses instead of mitigating them. These risks highlight the need for diversification in AI approaches and the integration of human oversight to ensure stability.

D. Accountability and Ethical Standards

When AI predictions are integrated into trading systems, establishing clear accountability is essential for ensuring regulatory compliance and mitigating risks. Organizations must determine who is responsible for decisions influenced by AI-generated outputs, particularly in scenarios where errors lead to financial losses. Moreover, ethical standards should be implemented to ensure fairness and transparency in AI systems. This includes safeguards against misuse and a commitment to prioritizing socially responsible applications of AI technology. Adherence to these principles is essential for fostering trust among stakeholders and minimizing potential harm.

E. Broader Implications

The ethical implications extend beyond financial gains or losses. Decisions made by AI models can affect market stability, investor confidence, and public trust in financial systems. It is the responsibility of developers and stakeholders to prioritize ethical AI practices and to foster collaboration between technologists and regulators.

VI. Future Work

While Quantum demonstrated strong predictive performance in comparison to human-designed strategies, significant opportunities exist to further enhance its capabilities. Expanding Quantum's data inputs to include alternative sources, such as financial forums, earnings call transcripts, and real-time retail sentiment, could provide a broader perspective on market dynamics and enrich its understanding of complex behaviors. Future iterations of Quantum could also leverage reinforcement learning techniques to dynamically adapt to changing market conditions, enhancing its ability to respond to unexpected events, such as black swan occurrences or policy shifts. To address the interpretability challenges inherent in AI-driven financial predictions, integrating explainable AI (XAI) methodologies, such as SHAP or LIME, would provide greater transparency into the model's decision-making process, fostering trust and enabling actionable insights. Additionally, incorporating real-time feedback loops into Quantum's architecture could facilitate continuous learning and refinement based on live market performance metrics, improving its predictive accuracy in high-frequency trading scenarios. Applying Quantum's architecture to other asset classes, such as commodities, cryptocurrencies, or emerging markets, would further test its robustness and adaptability across diverse financial environments. Furthermore, addressing potential systemic risks associated with AI-driven strategies, such as amplifying market volatility or herding behaviors, remains an essential area for future research, including exploring diversification techniques like model ensembles. Finally, prioritizing ethical AI development

by incorporating safeguards against biases and ensuring compliance with regulatory standards will enhance Quantum's societal impact and reliability in real-world applications.

VII. Conclusion

This study demonstrates that Quantum, a finance-specialized LLM, significantly outperforms both generalized LLMs (GPT-3, GPT-4) and other finance-focused models (FinGPT, FinBERT) in financial market prediction, particularly in areas of predictive accuracy, sentiment interpretation, and responsiveness to real-time data. Quantum's domain-specific training enabled it to capture subtle shifts in market sentiment and interpret complex financial language with a level of precision not achievable by generalized models. Our comparative analysis shows that Quantum achieved an accuracy rate of 78% for trend prediction and 82% for sentiment classification, both of which exceeded benchmarks set by FinGPT (72% and 76%, respectively) and FinBERT (74% and 78%) under identical testing conditions. However, Quantum's performance still trails human analysts in handling ambiguous language and in scenarios requiring deep contextual understanding, emphasizing the need for further model enhancements in those areas. These findings suggest that Quantum can serve as a powerful tool for augmenting financial analysis in data-intensive environments, though hybrid approaches that leverage both AI-driven insights and human expertise may offer the most reliable solutions for complex financial forecasting tasks. Future research will focus on expanding Quantum's adaptive capabilities, particularly in reducing contextual misinterpretations and enhancing its interpretative depth.

Acknowledgment

I would like to express my sincere gratitude to Professor Wang Hua for his invaluable guidance and support throughout this research. His expertise and insights have been instrumental in shaping the direction of this study, and it was his encouragement that inspired me to explore the application of large language models in financial prediction. This research would not have been possible without his mentorship, which has not only enriched my knowledge in the field but has also motivated me to pursue new ideas in quantitative finance.

VIII. References

[1] C. LIU, A. ARULAPPAN, R. NAHA, A. MAHANTI, J. KAMRUZZAMAN and A. I.-H. RA, "Large Language Models and Sentiment Analysis in Financial Markets: A Review, Datasets, and Case Study," *IEEE Access,* vol. 12, p. 134041–134056, 2024.

[2] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, G. Mai, N. Liu and T. Liu, "Revolutionizing Finance with LLMs: An Overview of Applications and Insights," *arXiv,* 2024.

[3] W. Luo1 and D. Gong, "Pre-trained Large Language Models for Financial Sentiment Analysis," *Journal of Financial Analytics,* p. 1–12, 2024.

[4] Y. Cao, Z. Chen, Q. Pei, F. Dimino, L. Ausiello, P. Kumar, K. Subbalakshmi and P. M. Ndiaye, "RiskLabs: Predicting Financial Risk Using LLMs Based on Multi-Sourced Data," *IEEE Transactions on Finance,* vol. vol. 17, p. 1120–1133, 2024.

[5] W. Zhang, Y. Deng, B. Liu, S. J. Pan and L. Bing, "Sentiment Analysis in the Era of Large Language Models: A Reality Check," *Journal of NLP Research,* p. 109–122, 2023.

[6] K. K. a and G. Germano, "Sentiment Trading with Large Language Models," *Finance Research Letters,* Vols. vol. 62, no. 105227, p. pp. 1–10, 2024.

[7] H. Zhang, F. Hua, C. Xu, H. Kong, R. Zuo and J. Guo, "Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements?," *Chinese Financial Analysis Journal,* p. pp. 55–67, 2024.

[8] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv,* 2019.

[9] T. Brown, B. Mann, N. Ryder and M. Subbiah, "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems," *Advances in Neural Information Processing Systems,* vol. vol. 33, p. 1877–1901, 2020.

[10] J. Delgadillo, J. Kinyua and C. Mutigwe, "FinSoSent: Advancing Financial Market Sentiment Analysis through Pretrained Large Language Models," *International Journal of Financial Data Science,* p. 78–89, 2024.

[11] S. Fatemi and Y. Hu, "A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning for Financial Sentiment Analysis," 2023.

[12] Y. Nie, Y. Kong, X. Dong, J. M. Mulvey and H. V. Poor, "A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges," 2024.