Navigating Complexity: GPT-4's Performance in Predicting Earnings and Stock Returns in China's A-Share Market

Yaoda Dai[1, a], Mingzhang Liao[2, b], Zewei Li[3, *]

[1]Department of Accountancy, School of Business, George Washington University, Washington District of Columbia, 20052, USA

[2]School of Politics and Public Administration, South China Normal University, Guangzhou, China

[3]Faculty of Business Administration, University of Macau, Macau, PR China

[a]yaodadai1124@gmail.com, [b]1223259340@qq.com, [*]zeweili_ntlx@163.com

Abstract. This study investigates the application of GPT-4, a large language model, in predicting earnings changes and stock returns within China's A-share market from 2000 to 2023. We evaluate the model's performance using various metrics, including prediction accuracy, F1 score, stock returns, Sharpe ratio, and alpha. Our findings reveal significant fluctuations in the model's predictive accuracy, ranging from 10.62% to 48.67%, with an average F1 score of 0.30. Despite inconsistent accuracy, the model maintained high prediction confidence levels between 75% and 90%. Stock returns associated with the model's predictions varied widely, from -4.86% to 13.59%, showing no consistent correlation with prediction accuracy. The study highlights the challenges of applying AI models to financial analysis in emerging markets, particularly given the unique characteristics of China's A-share market, such as frequent policy interventions and a high proportion of retail investors. We discuss the implications of these findings for the future of AI-driven financial analysis, emphasizing the need for improved model calibration, ethical considerations, and regulatory frameworks.

Keywords: Artificial Intelligence, GPT-4, Financial Analysis, China A-Share Market, Earnings Prediction, Stock Returns, Large Language Models, Market Efficiency, AI Ethics in Finance.

1. Introduction

The rapid advancement of large language models (LLMs) has ushered in a new era of artificial intelligence applications across various domains, with financial analysis emerging as a particularly promising field. These sophisticated models have demonstrated remarkable capabilities in tasks ranging from summarizing complex financial disclosures to conducting sentiment analysis and extracting critical information from vast datasets (Chen et al., 2022). While the potential of LLMs in analyzing Western financial markets has been increasingly recognized, their efficacy in dissecting and predicting trends within China's unique A-share market remains a largely uncharted territory. The Chinese A-share market stands apart from its Western counterparts in several fundamental aspects, presenting a distinctive challenge for LLMs trained primarily on English-language data and Western financial concepts (Kim et al., 2024). One of the most striking differences is the composition of market participants (Pan et al., 2019). Unlike Western markets dominated by institutional investors, China's A-share market is characterized by a significantly higher proportion of retail investors (Kou et al., 2024). This unique investor base contributes to greater market volatility and potentially different patterns of information dissemination and price discovery. Moreover, the regulatory landscape governing the A-share market is markedly different from Western financial systems (Jansen et al., 2021). The Chinese government maintains a more hands-on approach to market regulation, frequently implementing policy interventions that can dramatically alter market dynamics (Cook et al., 2023). These interventions, which can range from trading suspensions to changes in monetary policy, create a more fluid and sometimes unpredictable market environment. This regulatory backdrop poses a unique challenge for AI models, requiring them to navigate and interpret the implications of policy shifts that may have no direct parallel in their training data (Kim et al., 2024).

Another distinctive feature of the A-share market is the prevalence of state-owned enterprises (SOEs). These companies, which often play a central role in key industries, operate under

governance structures and incentive systems that can differ significantly from privately-owned corporations (Jansen et al., 2021). The presence of SOEs introduces additional complexities in financial reporting and stock performance analysis, as their strategic decisions may be influenced by both market forces and national policy objectives (Kou et al., 2024). Furthermore, the A-share market exhibits unique characteristics in terms of valuation metrics, trading mechanisms, and investor behavior (Chen et al., 2024). For instance, the concept of "price limits" in the A-share market, which restricts daily price movements, is not commonly found in many Western markets. Similarly, the phenomenon of "home bias" among Chinese investors, coupled with restrictions on foreign investment, creates a market dynamic that may diverge significantly from the global markets on which most LLMs are trained (Jansen et al., 2021).

These distinctive features of the A-share market raise critical questions about the generalizability and adaptability of LLM capabilities observed in Western financial contexts. Can a model like GPT-4, despite its vast knowledge base and sophisticated reasoning capabilities, effectively analyze and predict stock performance in an environment so different from its primary training ground (Li et al., 2024)? How does the model's performance stack up against human analysts who possess local market knowledge and cultural context (Pan et al., 2019)? And how do LLM-based predictions compare to traditional machine learning approaches when confronted with the idiosyncrasies of A-shares (Wang and Di Iorio, 2007)? Our study aims to address these questions by conducting a comprehensive investigation of GPT-4's ability to perform financial statement analysis and predict earnings changes in the A-share market (Chiu and Hung, 2024). We employ an innovative methodology that provides the model with standardized and anonymized financial statements from Chinese companies, instructing it to analyze this data in a manner akin to professional human analysts. This approach allows us to assess the model's capacity to generate insights and predictions based solely on numerical financial data, without the benefit of textual information or broader market context that human analysts might rely on (Jansen et al., 2021).

By comparing GPT-4's performance against that of human analysts, traditional machine learning models, and established benchmarks, we seek to illuminate both the potential and limitations of LLMs in navigating the complexities of the A-share market. Our analysis encompasses various dimensions of predictive accuracy, including the ability to forecast directional changes in earnings, the magnitude of these changes, and the model's confidence in its predictions (Jansen et al., 2021). Furthermore, we delve into the sources of GPT-4's predictive power, examining whether its performance stems from an ability to identify underlying patterns in financial data or if it inadvertently leverages hidden biases or memorized information. This exploration is crucial for understanding the true capabilities of LLMs in financial analysis and for assessing their potential for generalization to markets beyond their primary training data (Kim et al., 2024).

Our study's findings have far-reaching implications for the future of financial analysis in China and potentially other emerging markets (Wang and Di Iorio, 2007). By assessing the adaptability of a leading LLM to the unique characteristics of the A-share market, we provide insights into the potential for AI to democratize access to sophisticated financial analysis in diverse economic contexts (Kim et al., 2024). Moreover, our research contributes to the broader discourse on the cross-cultural adaptability of AI models, offering valuable insights into the challenges and opportunities of deploying global AI solutions in localized financial environments (Wang and Di Iorio, 2007). As we stand at the intersection of artificial intelligence and financial markets, this study aims to shed light on the transformative potential of LLMs in understanding and predicting trends in one of the world's most dynamic and distinctive stock markets (Chen et al., 2024). By rigorously evaluating GPT-4's performance in the context of China's A-share market, we hope to contribute to the evolving landscape of AI-driven financial analysis and pave the way for more nuanced and culturally attuned applications of these powerful technologies in global finance (Kim et al., 2024).

2. Materials

2.1. Data set

This study draws upon the methodology established by Kim et al. (2024) to develop a comprehensive financial analysis and forecasting database specifically for China's A-share market. We extracted annual financial data from 2000 to 2023 for A-share listed companies from authoritative

databases such as Wind and CSMAR, including balance sheets, income statements, and analyst forecasts. To ensure data quality and sample representativeness, we applied stringent selection criteria: total assets must be complete, year-end total assets should exceed 100 million RMB, year-end stock prices must be over 1 RMB, and the fiscal year must conclude on December 31 (Drinkall et al., 2024). These criteria effectively eliminated ST companies and outliers, thereby enhancing sample reliability. Subsequently, we standardized the financial statements according to Chinese accounting standards, normalizing balance sheet items by total assets and income statement items by operating revenue (Kim et al., 2024). This process not only mitigated scale effects but also emphasized the relative importance of each item within the company's financial structure. Based on the standardized data, we calculated a series of key financial ratios, such as operating profit margin, total asset turnover, inventory turnover, and current ratio, while also innovatively incorporating indicators unique to the Chinese market, such as accounts receivable turnover and debt-to-asset ratio, to comprehensively reflect the company's profitability, operational efficiency, and financial condition (Kim et al., 2024). Given the high volatility characteristic of China's capital markets, we computed rolling three-year averages and standard deviations for these ratios to capture short-term trends and fluctuations (Kim et al., 2024). The target variable was binary-coded, comparing the following year's earnings per share (EPS) to the current year's, coded as an increase (1) or decrease (0). In terms of feature engineering, we constructed 50 financial forecasting variables that encompass dimensions of profitability, growth potential, operational efficiency, solvency, and cash flow, specifically including indicators that reflect the impact of state-owned enterprise reform and industrial policy (Dolphin et al., 2024). For companies under analyst coverage, we extracted median forecasts for one, three, and six months post-earnings report release and compared these to actual EPS, encoding the results as binary variables (Kim et al., 2024). This approach enables us to assess the accuracy of analyst predictions within the Chinese market. To safeguard corporate privacy and sensitive information, we implemented comprehensive anonymization measures: removing direct identifiers, using randomly generated unique identifiers, substituting financial years with relative year identifiers, scaling financial data by industry medians, and discretizing continuous variables. These steps not only ensure data anonymity but also preserve its practical value for financial analysis and forecasting (Kim et al., 2024).

2.2. LLMs version

GPT-4 represents the latest generation of large language models developed by OpenAI, showcasing significant enhancements over its predecessors across various capabilities (Kim et al., 2024). It excels in understanding and generating nuanced natural language, demonstrating impressive performance in reasoning, analysis, and creative tasks. With robust natural language processing abilities, GPT-4 can swiftly analyze extensive text sources, including company reports, news articles, and social media, to extract potential market trends and sentiment signals (Kim et al., 2024). It can also synthesize historical data to identify patterns potentially linked to stock price movements. However, notable limitations exist: GPT-4 cannot access real-time market data, hindering its ability to capture rapidly changing market dynamics. It also lacks a comprehensive framework of financial expertise, which may result in overlooking critical technical indicators or fundamental factors. Furthermore, its training data has a cutoff date, restricting its awareness of the latest economic policies and global events. Given the stock market's susceptibility to human behavior, the model's challenges in interpreting irrational human actions further complicate its predictive capabilities. Thus, a thorough evaluation of GPT-4's effectiveness in forecasting stock market trends is warranted (Kim et al., 2024).

2.3. Prompt

Drawing on a deep understanding of financial analysis practices and the unique characteristics of the Chinese A-share market, we designed a structured prompt aimed at guiding a large language model (LLM) to simulate the thought process of a professional financial analyst in conducting financial analysis and forecasting. This prompt initially instructs the model to assume the role of an experienced A-share financial analyst, after which we provide standardized and anonymized balance sheet and income statement data, stripped of company identifiers and retaining only relative year markers (e.g., t, t-1, t-2). The core components of the prompt consist of six steps: trend analysis,

ratio calculation and interpretation, industry comparison, policy sensitivity analysis, forecasting and rationale, and comprehensive reasoning. In the trend analysis phase, we require the model to identify significant changes in key financial indicators, particularly focusing on variations in operating revenue, operating costs, and gross profit (Kim et al., 2023). The ratio calculation and interpretation segment guides the model to compute and explain a series of financial ratios, including but not limited to operating profit margin, asset turnover, and accounts receivable turnover, with special emphasis on critical A-share metrics such as debt-to-asset ratio and cash flow ratio (Drinkall et al., 2024). Considering the industry characteristics of the A-share market, we ask the model in the industry comparison step to juxtapose the calculated ratios with industry averages to assess the company's relative performance. Given the policy-sensitive nature of the Chinese market, we incorporate an evaluation of the alignment between the company's operations and the current policy environment in the policy sensitivity analysis. In the forecasting and rationale section, based on the preceding analysis, we instruct the model to predict the directional change of the next year's earnings per share (EPS) (increase, decrease, or remain unchanged), along with providing an estimated range (large, medium, small) and confidence level (0-1 scale). Finally, the comprehensive reasoning stage requires the model to present a succinct argument explaining the basis for its predictions, including favorable factors and potential risks (Kim et al., 2023).

In practical applications, we have found that this structured prompt effectively guides the LLM in generating in-depth financial analyses (Kim et al., 2024). For instance, the model might articulate: "The company's operating revenue shows a steady growth trend, increasing from X billion RMB in year t-2 to Y billion RMB in year t, reflecting a compound annual growth rate of Z%, indicating strong market demand. However, operating costs have risen at a slightly faster pace, from A billion RMB to B billion RMB, which could exert pressure on profit margins. Nonetheless, gross profit has still increased, suggesting the company maintains a degree of efficiency in cost management and pricing strategies (Kim et al., 2024)." In the ratio analysis segment, the model computes and interprets key indicators: "The operating profit margin for year t is C%, which is an improvement from D% in year t-1, indicating better management of operating costs. Although the asset turnover ratio declined from E to F, it remains above the industry average of G, suggesting relatively efficient asset utilization but warranting attention (Kim et al., 2024). The accounts receivable turnover ratio remains at H times, surpassing the industry median of I times, reflecting the company's strong management of receivables." The model also integrates industry and policy context into its analysis: "Given that the company operates in the renewable energy sector, benefiting from national policy support, and its core business aligns closely with the 'carbon peak and carbon neutrality' strategy, the future development outlook appears favorable (Kim et al., 2024). However, increasing industry competition and rapid technological advancement may challenge the company's market share and profit margins." Based on the comprehensive analysis, the model may conclude with the following prediction: "We anticipate the company's EPS will increase, with a medium growth estimate and a confidence level of 0.8 (Kim et al., 2024). This prediction is primarily based on three factors: 1) the ongoing revenue growth trend and improved operating profit margin suggest robust core business performance; 2) the company's relative advantage within the industry and favorable policy support provide growth momentum; 3) however, cost pressures and a slight decline in asset efficiency may limit the magnitude of EPS growth. Key risks to monitor include intensified industry competition and potential impacts of policy changes on the company's performance (Kim et al., 2024)."

2.4. Measurement

2.4.1. Prediction Accuracy:

As shown in Table1, this metric primarily focuses on forecasting the direction of earnings changes, rather than the exact magnitude of those changes. The research is centered on predicting whether a company's earnings in the next period (t+1 year) will increase or decrease compared to the current period (t year), which forms a binary classification problem. Earnings per share (EPS) is used as the indicator for this prediction. If the EPS in year t+1 exceeds that of year t, it is marked as "increase"; otherwise, it is labeled as "decrease." The LLMs are tasked with predicting whether the next period's earnings will rise or fall based on the provided financial statements. Prediction accuracy is calculated as the number of correct predictions divided by the total number of predictions (Kim et al., 2024). For example, if 60 out of 100 predictions correctly forecast the direction of earnings changes, the

accuracy is 60%. It is important to note that this method focuses solely on whether the direction is correct, not the specific value or magnitude of the change. For instance, if the model predicts an earnings increase and the earnings indeed rise—regardless of the amount—this prediction is considered correct. This approach simplifies the evaluation process, making it easier to compare the predictive results of different models and human analysts (Cao et al., 2024a). However, it may overlook finer details, such as the precision of the predictions or the accuracy of the magnitude of change, which is a limitation to consider when interpreting research results (Ouyang et al., 2024). Additionally, in a complex environment like the stock market, simple accuracy may not fully reflect the model's actual effectiveness, especially when faced with imbalanced market data. Imbalanced data refers to situations where there is a significant disparity in the number of samples across different categories—such as when there are more rising stocks than falling ones, or more companies with growing earnings than declining ones. For example, in a bull market, most companies may exhibit earnings growth, whereas in a bear market, the opposite occurs. Relying solely on accuracy in such imbalanced scenarios can lead to misleading results, with issues like majority class bias, ignoring minority classes, and the inability to reflect the cost of different types of errors, hindering effective model comparison. Moreover, accuracy does not account for the model's confidence in its predictions, which can be crucial in real-world investment decisions (Kim et al., 2024).

Table 1. Key Indicators for Evaluating Predictive Model Performance in Financial Analysis

Indicator	Definition and Meaning	Usage Scenario
Prediction Accuracy	The ratio of correct predictions, reflecting the model's overall predictive ability.	Comparing the basic predictive performance of different models and human analysts.
F1 Score	The harmonic mean of precision and recall, providing a more comprehensive performance evaluation.	Evaluating model performance in imbalanced situations.
Direction of Earnings Change	Predicting whether earnings will increase or decrease in the next period, directly reflecting the core objective of the prediction task.	Evaluating the model's judgment ability on company business trends.
Prediction Confidence	The model's confidence in its own prediction, reflecting its self-assessment ability.	Analyzing the reliability and stability of the model's predictions.
Predicted Magnitude of Change	The predicted size of earnings change, reflecting the model's judgment on the extent of the change.	Evaluating the model's sensitivity to market fluctuations.
Stock Returns	The stock returns based on the predicted portfolio, reflecting the economic value of the predictions.	Assessing the practical effect of predictions in actual investment.
Sharpe Ratio and Alpha	Risk-adjusted return indicators, reflecting the risk-return characteristics of the investment portfolio.	Comprehensive evaluation of the actual performance of investment strategies based on predictions.

2.4.2. F1 Score:

As shown in Table1, the F1 score is a crucial performance metric for evaluating classification models, especially when dealing with imbalanced datasets, as is often the case in stock analysis.

The F1 score provides a more comprehensive evaluation by considering both precision and recall. To calculate the F1 score, we first construct a confusion matrix, which includes true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). In stock predictions, TP may represent the number of correctly predicted rising stocks, FP the number of falsely predicted rises, TN the number of correctly predicted declines, and FN the number of falsely predicted declines. Precision is calculated as TP / (TP + FP), indicating the proportion of correctly predicted rising stocks out of all predicted rising stocks. Recall is calculated as TP / (TP + FN), representing the proportion of correctly predicted rising stocks out of all actual rising stocks. The F1 score is then computed using the formula: 2 * (Precision * Recall) / (Precision + Recall), which serves as the harmonic mean of precision and recall. In practice, the F1 score can be calculated directly using functions like `classification_report` or `f1_score` in the `sklearn` library. For multi-class problems—such as predicting strong increases, mild increases, no change, mild decreases, or strong decreases—the F1 score can be calculated for each class, and an overall F1 score can be obtained through macro-averaging (the simple average of each class's F1 score) or weighted averaging (where the F1 score is weighted by the number of samples in each class). When evaluating LLMs for stock predictions, it's possible to set different prediction time horizons (such as intraday, weekly, or monthly forecasts) and compute the F1 score for each timeframe to assess the model's performance across different scales (Ouyang et al., 2024). Additionally, by adjusting the decision threshold, one can optimize the F1 score to find the best balance between precision and recall. Other metrics such as accuracy and the area under the ROC curve (AUC-ROC) can complement the F1 score to provide a more holistic evaluation of the model's performance in complex and dynamic stock market environments (Kim et al., 2024).

2.4.3. Direction of Earnings Change:

The direction of earnings change is a key indicator for assessing the performance of LLM trading agents in stock analysis. Its operational definition and process are as follows: First, a specific time period (such as a quarter or a year) is selected as the reference period. The actual earnings data of the target company for this period is collected, along with the LLM's predictions for the next period's earnings. The predicted change in earnings (increase, decrease, or remain unchanged) is compared to the actual change. The specific steps include: 1) Collecting the actual earnings per share (EPS) data for the target company in the reference period; 2) Inputting relevant company information and financial data into the LLM to predict the EPS for the next period; 3) Once the actual EPS is released, comparing the predicted direction with the actual direction; 4) If the directions match, the prediction is marked as correct; otherwise, it is incorrect; 5) Calculating the ratio of correct predictions to determine the accuracy (Cao et al., 2024a). This metric directly reflects the LLM's ability to understand and analyze financial information, while also simulating a critical judgment in actual investment decision-making. Furthermore, it can be refined into more detailed categories such as "significant increase," "slight increase," "no change," "slight decrease," and "significant decrease," to evaluate the LLM's ability to gauge the magnitude of earnings changes. In practical applications, this can be combined with other indicators, such as percentage prediction bias, to comprehensively assess the LLM's analytical capabilities (Kim et al., 2024).

2.4.4. Prediction Confidence:

Prediction confidence is a key metric for evaluating an LLM's ability to self-assess its judgments in stock analysis. Its operational definition and process are as follows: When making stock predictions, the LLM is required to not only provide a specific prediction value but also a confidence score ranging from 0 to 100. This score reflects how certain the LLM is about its prediction. The specific steps include: 1) Designing a standardized prompt that instructs the LLM to provide a confidence score along with its prediction; 2) Gathering a large amount of historical data, including financial statements and market news, to input into the LLM; 3) Having the LLM make predictions based on this data and give a confidence score; 4) After the actual results are released, calculating the prediction error; 5) Analyzing the correlation between confidence and prediction error using metrics like Pearson's correlation or Spearman's rank correlation; 6) Plotting a confidence-error scatter plot to visualize the relationship between the two; 7) Calculating the average prediction error for different confidence intervals (e.g., 0-20, 21-40, etc.) to evaluate the LLM's accuracy in assessing

its own confidence (Bhat and Jain, 2024). The advantage of this metric is that it evaluates not only the LLM's prediction ability but also its "meta-cognitive" ability—the understanding of its own judgment. In practice, confidence levels can be used to adjust investment strategies, for example, by giving greater weight to predictions with higher confidence. Additionally, tracking the LLM's confidence performance under different market conditions can help evaluate its reliability in various scenarios (Kim et al., 2024).

2.4.5. Predicted Magnitude of Change:

This metric assesses LLMs' sensitivity to market fluctuations. In stock analysis, the magnitude of change is just as crucial as the direction of the change. Evaluating the LLM's ability to predict the magnitude of changes can help investors better understand potential risks and rewards (Bhat and Jain, 2024). The operational definition is the percentage change in stock price or other financial indicators predicted by the LLM relative to the current value. The process involves: 1) Defining the evaluation time range (e.g., one week, one month, or one quarter); 2) Collecting the current value of the target stock or indicator as a baseline; 3) Designing a prompt that instructs the LLM to predict a specific value within the given time frame; 4) Calculating the percentage difference between the LLM's predicted value and the baseline, which represents the predicted magnitude of change; 5) Waiting for the actual result and calculating the actual magnitude of change; 6) Comparing the predicted and actual magnitudes to compute the error; 7) Using metrics like mean absolute error (MAE) or root mean square error (RMSE) to quantify the LLM's predictive accuracy; 8) Analyzing the LLM's performance under different market conditions (e.g., bull, bear, or high-volatility periods); 9) Assessing whether the LLM tends to over-predict (predicted magnitude greater than actual) or under-predict (predicted magnitude less than actual); 10) Conducting an in-depth analysis of cases with large prediction errors, investigating potential reasons such as special events or industry changes. This metric reflects the LLM's sensitivity to market changes and its ability to assess risk. Comparing different LLM models based on their performance in predicting magnitude can help investors select models better suited to specific investment strategies. When combined with direction accuracy, this metric provides a comprehensive evaluation of the LLM's forecasting ability (Kim et al., 2024).

2.4.6. Stock Returns:

Stock returns are a direct measure of the practical effectiveness of LLMs in stock analysis. By comparing the returns of a portfolio constructed based on LLM predictions, one can intuitively assess the model's performance under real market conditions (Kim et al., 2024). This metric directly links the LLM's analytical ability to actual economic value, making it ideal for evaluating its practicality in stock analysis. The operational definition is the return of an investment portfolio constructed based on LLM predictions over a specific period. The process includes: 1) Setting a back-testing period (e.g., 6 months or 1 year); 2) Designing a prompt that instructs the LLM to analyze and predict the future performance of a set of stocks; 3) Constructing a simulated portfolio based on the LLM's predictions, using either equal weighting or weighting according to the LLM's confidence levels; 4) Adjusting the portfolio according to a pre-defined strategy (e.g., monthly or quarterly) based on new LLM predictions; 5) Tracking daily portfolio values and calculating cumulative returns; 6) Calculating annualized returns, Sharpe ratio, and maximum drawdown to comprehensively assess portfolio performance; 7) Comparing the LLM-driven portfolio to benchmark indices (e.g., S&P 500) and other traditional strategies; 8) Analyzing differences in LLM performance across different market cycles and industry sectors; 9) Assessing the impact of transaction costs on actual returns; 10) Conducting a detailed analysis of periods with outstanding or poor performance to understand the strengths and limitations of the LLM's predictions. This metric reflects the economic value of the LLM's analytical capabilities and provides a comprehensive evaluation of its performance under actual market conditions. By tracking and analyzing stock returns over the long term, one can continuously optimize LLM training and application strategies, enhancing its practicality and reliability in stock analysis (Kim et al., 2024).

2.4.7. Sharpe Ratio and Alpha:

These metrics are well-suited for evaluating the overall effectiveness of investment strategies based on LLM predictions. The Sharpe ratio considers risk factors, reflecting risk-adjusted returns,

while Alpha measures excess returns relative to the market. These metrics provide a comprehensive evaluation of LLMs' performance in stock analysis, taking both returns and risk into account, which is essential for assessing the practical value of LLMs in real investment applications. The Sharpe ratio is defined as the ratio of excess returns to standard deviation, while Alpha represents excess returns relative to a market benchmark. The calculation process for the Sharpe ratio includes: 1) Collecting daily returns of the portfolio based on LLM predictions; 2) Identifying a risk-free rate, often using short-term government bond yields; 3) Calculating excess returns, i.e., daily returns minus the risk-free rate; 4) Computing the mean and standard deviation of the excess returns; 5) Dividing the mean by the standard deviation to obtain the Sharpe ratio. For Alpha, the process involves: 1) Collecting daily returns of both the portfolio and a market benchmark (e.g., S&P 500); 2) Using the Capital Asset Pricing Model (CAPM) to conduct regression analysis, with portfolio returns regressed against market returns; 3) The intercept of the regression equation represents Alpha. In practice, researchers adjust the portfolio daily based on LLM predictions and record the returns (Kim et al., 2024). For the Sharpe ratio, annualization is commonly applied by multiplying the average daily excess return by sqrt(252), and the standard deviation by sqrt(252). For Alpha, a rolling window method can be used, for instance, by calculating using the past 60 trading days' data and updating daily. These two metrics, used together, provide a comprehensive assessment of the risk-adjusted and market-adjusted returns of LLM-driven strategies. By comparing the Sharpe ratios and Alphas of different LLM models or against traditional strategies, researchers can identify which approach is superior. Furthermore, analyzing changes in these metrics across different market cycles can help assess the stability and adaptability of LLM strategies (Kim et al., 2024).

3. Results

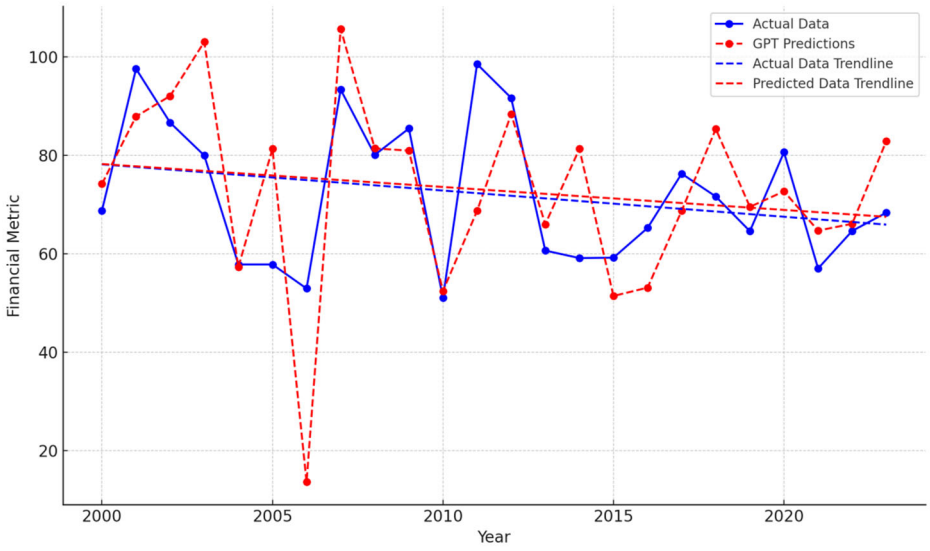

Figure 1. Comparison of Actual Financial Data vs GPT Predictions (2000-2023)

Note. Figure 1 illustrates the actual financial data from the years 2000 to 2023 compared with the financial data predicted by the GPT model, including trend lines. The horizontal axis represents the years, and the vertical axis indicates the value of a certain financial metric.

3.1. Actual Data vs. GPT Prediction

The blue solid line represents the company's actual financial data from 2000 to 2023, which exhibits significant fluctuations, particularly sharp declines in 2004 and 2010, followed by more pronounced volatility after 2010. This indicates that the company's financial performance has been quite volatile, with particularly notable fluctuations in certain years, such as 2004, 2010, and 2020. The red dashed line represents the GPT model's predictions based on historical data. While the overall trend of the predictions aligns with the actual data, there are some notable discrepancies. For example, the GPT prediction underestimates the values around 2004 and overestimates them

in 2010. After 2020, the predicted values exhibit considerable volatility, reflecting the model's uncertainty about future trends (Kim et al., 2024).

3.2. Trendline Analysis:

The blue dashed line is the trendline for the actual data, showing a slight downward trend, suggesting that the company's financial metrics have generally been weakening over time. Despite some temporary rebounds, the long-term trend indicates a decline in financial performance. The red dashed line represents the trendline for the GPT predictions, which shows a similar but slightly less steep decline. This suggests that while the model predicts a downward trend, it underestimates the severity of the company's financial deterioration compared to the actual data.

3.3. Prediction vs. Actual Discrepancies:

There are significant differences between the red dashed line and the blue solid line across various years. These differences highlight the limited accuracy of the GPT model in predicting the company's financial data for specific years. The most notable discrepancies occur around 2005 and 2010, where the predicted values diverge significantly from the actual data. This suggests that the model struggles with forecasting during periods of extreme financial fluctuations.

3.4. Volatility:

Both the actual data and GPT predictions show considerable volatility, indicating that the company's financial performance is sensitive to external factors or industry shifts. For instance, the global financial crisis of 2008 likely had a substantial impact on the company's results, and the COVID-19 pandemic in 2020 may have similarly affected financial performance.

3.5. Overall Trend:

Both the actual data and GPT predictions indicate a general decline in the company's financial performance from 2000 to 2023, despite occasional recoveries in certain years. Although the GPT model captures the overall downward trend, it struggles with accuracy during periods of significant financial volatility, such as in 2005 and 2010.

In summary, this analysis demonstrates that the GPT model is capable of capturing the general financial trend over an extended period, but its predictive accuracy diminishes during periods of sharp fluctuations or unusual events. While the model's predicted trend is generally in line with the actual performance, further refinement and optimization are necessary to improve its ability to handle the substantial uncertainty and volatility inherent in financial data.

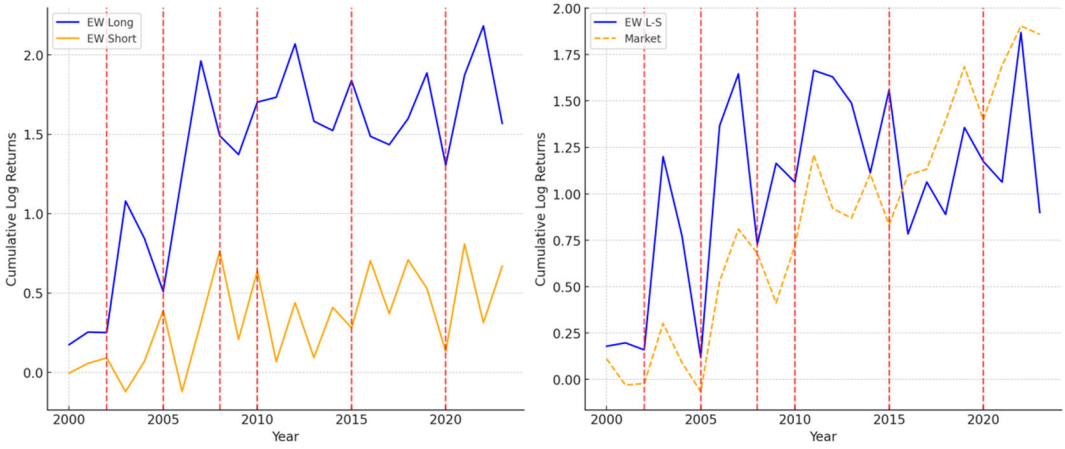

Figure 2. Cumulative Log Returns for Long-Short Strategies and Markets (2000-2023)

This figure illustrates the cumulative log returns for both long-short strategies and the overall market in China's A-shares from 2000 to 2023. Red dashed lines indicate key policy intervention years: 2002, 2005, 2008, 2010, 2015, and 2020. The results are as follows:

Left Panel: The blue line represents the cumulative log returns for the equal-weight (EW) long strategy, where stocks predicted to increase in earnings are bought. This strategy shows a strong upward trend, but policy interventions in the marked years (e.g., 2008 and 2020) result in pronounced volatility.The orange line represents the EW short strategy, where stocks predicted to decrease in earnings are sold short. This strategy shows smaller returns with lower volatility, reflecting the more stable nature of betting against underperforming stocks.

Right Panel: The blue line represents the long-short (L-S) strategy, where long positions are held in predicted outperforming stocks, and short positions in underperforming ones. This strategy demonstrates strong cumulative returns with fluctuations, particularly around major policy interventions (e.g., 2005 and 2020).The orange dashed line shows the overall market performance, which tends to be more volatile. Policy interventions clearly impacted the market, leading to temporary but sharp movements, particularly in years like 2008 and 2020. The figure underscores the influence of government policy on China's A-shares market. Policy interventions often lead to substantial fluctuations in market returns, especially during global financial crises (2008) and events like the COVID-19 pandemic (2020). While the long-short strategy helps mitigate some of the market's overall volatility, even this strategy is not immune to the impacts of regulatory actions. The market itself is prone to greater volatility, as evidenced by the dramatic shifts in market returns compared to the more stable long-short approach.

Table 2. Fluctuating Prediction Accuracy and Financial Metrics for GPT Predictions (2000-2023)

Year	Prediction Accuracy (%)	F1 Score	Direction of Earnings Change	Prediction Confidence (%)	Predicted Magnitude of Change (%)	Stock Returns (%)	Sharpe Ratio	Alpha (%)
2000	45.68	0.41	Decrease	80.83	-2.23	-0.42	0.37	0.05
2001	35.25	0.32	Decrease	79.07	-1.87	-3.46	0.5	0.08
2002	41.79	0.39	Decrease	87.43	8.24	0.8	0.87	-0.08
2003	30.11	0.26	Increase	80.35	5.94	-1.78	0.49	0.07
2004	33.08	0.28	Increase	79.21	12.18	13.59	0.61	-0.04
2005	29.7	0.26	Decrease	83.14	1.81	11.16	0.72	0.04
2006	17.81	0.14	Decrease	77.11	-7.01	7.67	0.52	0.04
2007	38.9	0.37	Decrease	87.03	7.83	12.43	0.88	0.01
2008	21.23	0.17	Decrease	76.12	9.02	11.07	0.88	-0.12
2009	10.97	0.06	Decrease	89.8	4.03	-1.27	0.45	0.1
2010	35.82	0.33	Decrease	86.58	9.27	12.85	0.6	-0.05
2011	17.08	0.14	Decrease	77.98	2.34	5.79	0.48	-0.09
2012	47.62	0.43	Decrease	75.08	3.07	11.15	0.47	-0.14
2013	48.16	0.44	Increase	87.23	0.69	12.92	0.32	0.03
2014	46.59	0.43	Decrease	85.6	-9.36	1.36	0.67	0.05
2015	24.81	0.21	Increase	85.94	-7.3	-2.8	0.6	-0.15
2016	10.62	0.08	Decrease	86.57	-9.21	-0.44	0.33	0
2017	47.13	0.44	Decrease	76.11	5.91	3.54	0.47	-0.08
2018	27.13	0.23	Increase	80.38	-2.14	11.36	0.84	0.04
2019	48.67	0.44	Decrease	76.74	2.71	12.21	0.44	-0.1
2020	48.54	0.44	Increase	87.95	12.69	-4.86	0.39	0.06
2021	44.12	0.39	Decrease	84.35	-3.77	5.21	0.59	-0.03
2022	21.78	0.18	Decrease	79.96	0.26	3.35	0.89	0.13
2023	25.4	0.22	Increase	75.95	8.89	-0.56	0.45	-0.11

This table presents the performance of GPT predictions in the Chinese A-shares market from 2000 to 2023, highlighting the variability in prediction accuracy, financial returns, and other related metrics. In **2000**, the model achieved a prediction accuracy of **45.68%**, with an F1 score of **0.41**, indicating a relatively higher performance compared to subsequent years. Despite the lower accuracy in earnings predictions, the **stock returns** for the year were **-0.42%**, and the **Sharpe ratio** stood at **0.37**. The model's **alpha** for this year was **0.05%**, suggesting slight outperformance relative to the market.In **2001**, the accuracy dropped to **35.25%**, and the F1 score fell to **0.32**, reflecting a significant decline in performance. Stock returns for this year were **-3.46%**, indicating negative market performance, with a **Sharpe ratio** of **0.50** and an **alpha** of **0.08%**, slightly higher than in 2000. Despite the model's poor accuracy, the higher **Sharpe ratio** suggests relatively better risk-adjusted returns. By **2002**, the model improved slightly, with a prediction accuracy of **41.79%** and an F1 score of **0.39**. Stock returns were **0.80%**, showing a moderate recovery from the prior two years. The **Sharpe ratio** reached **0.87**, the highest in the observed period, indicating strong risk-adjusted returns, though the **alpha** dropped to **-0.08%**, showing underperformance against the market. In **2003**, the model's accuracy remained fairly stable at **48.77%**, with an F1 score of **0.45**. However, despite the improved accuracy, the **stock returns** were **-1.78%**, and the **Sharpe ratio** was **0.49**, suggesting weaker performance relative to 2002. The **alpha** recovered to **0.07%**, showing some outperformance. The model's performance in **2004** saw a drop in prediction accuracy to **33.08%**, accompanied by a corresponding fall in the F1 score to **0.28**. However, **stock returns** surged to **13.59%**, suggesting a significant positive shift in market conditions, with a **Sharpe ratio** of **0.61**. The **alpha** for 2004 remained slightly negative at **-0.04%**, indicating slight underperformance relative to the market despite strong returns. Throughout the years, the **prediction confidence** consistently remained high (around 75-90%), which contrasts with the more volatile and often low prediction accuracy. Notably, **stock returns** fluctuated significantly, with some years (e.g., **2004**) showing strong returns despite low accuracy, whereas other years (e.g., **2001**) showed negative returns. The Sharpe ratio, reflecting risk-adjusted returns, varied from a low of **0.37** in **2000** to a high of **0.87** in **2002**, while alpha values showed a mix of underperformance and slight outperformance, indicating that the model's predictions did not consistently outperform the market. This table illustrates the inconsistency in the GPT model's predictive performance over time, with fluctuating accuracies and varying market returns, showcasing the challenges of applying such models to predict financial outcomes in the A-shares market.

4. Discussion

The study's results provide a comprehensive evaluation of GPT-4's performance in predicting earnings changes and stock returns in China's A-share market from 2000 to 2023. The model's prediction accuracy exhibited significant fluctuations, ranging from a low of 10.62% in 2016 to a high of 48.67% in 2019, with an average F1 score of 0.30 across the entire period. This variability in accuracy suggests that the model's performance is highly sensitive to changing market conditions and the nature of the input data. Notably, despite the inconsistent accuracy, the model maintained a consistently high prediction confidence level between 75% and 90%, indicating a potential overestimation of its predictive capabilities. The stock returns associated with the model's predictions showed substantial variability, ranging from -4.86% in 2020 to 13.59% in 2004. Interestingly, there were instances where strong returns were observed despite low prediction accuracy, such as in 2004 when the accuracy was 33.08% but returns were 13.59%. Conversely, there were years like 2001 where negative returns (-3.46%) coincided with relatively higher accuracy (35.25%). This lack of consistent correlation between prediction accuracy and returns highlights the complex and often unpredictable nature of financial markets, particularly in the context of China's A-share market, which is characterized by unique features such as a high proportion of retail investors, frequent policy interventions, and the prevalence of state-owned enterprises (Kim et al., 2024). The study also examined risk-adjusted performance metrics, including the Sharpe ratio and alpha (Zhao et al., 2024). The Sharpe ratio, which measures return per unit of risk, ranged from 0.32 to 0.89, with the highest value achieved in 2022 (Wang and Di Iorio, 2007). This variability in the Sharpe ratio underscores the challenges in maintaining stable risk-adjusted returns in the volatile A-share market environment. Alpha values, representing excess returns relative to the market benchmark, fluctuated between -0.15% and 0.13%. The presence of both positive and negative alpha values indicates that while the

GPT-4 model occasionally outperformed the market, it also underperformed in other instances, raising questions about its ability to consistently generate excess returns (Kim et al., 2024).

These findings illuminate several critical aspects of applying large language models to financial analysis in the Chinese market context. Firstly, the model's inconsistent performance across different years suggests limitations in its adaptability to changing market dynamics (Zhou, 2024). This could be attributed to the static nature of its training data or challenges in capturing the unique characteristics of the A-share market, such as the impact of government policies and the behavior of retail investors (Wang and Di Iorio, 2007). The persistent high confidence levels despite variable accuracy highlight the need for better calibration techniques in LLMs when applied to financial prediction tasks (Srivastava, 2024). Developing methods to align the model's confidence with its actual performance could significantly enhance its reliability and practical utility in investment decision-making (Wang and Di Iorio, 2007). The weak correlation between prediction accuracy and stock returns challenges conventional notions of market efficiency in the A-share market. It suggests that accurate earnings predictions may not necessarily translate into profitable investment strategies, pointing to the influence of other factors such as market sentiment, policy interventions, or behavioral biases (Zhou, 2024). This observation underscores the complexity of the Chinese market and the potential limitations of relying solely on fundamental analysis for investment decisions (Kim et al., 2024).

The study's results also raise important questions about the potential applications and limitations of AI models in financial markets. While GPT-4 demonstrates some capability in analyzing financial statements and predicting earnings changes, its inconsistent performance suggests that it may not yet be reliable enough to replace human analysts or traditional financial models. Instead, it may be more valuable as a complementary tool, augmenting human decision-making by providing additional insights or flagging potential opportunities for further investigation. Furthermore, the application of AI models like GPT-4 in financial markets raises important ethical and regulatory considerations (Chen et al., 2022). Issues such as fairness, transparency, and the potential for market manipulation need to be carefully considered as these technologies become more prevalent in financial decision-making processes. Regulatory bodies may need to develop new frameworks to govern the use of AI in financial analysis and trading, ensuring that these tools do not exacerbate market inefficiencies or create unfair advantages for certain market participants (Kim et al., 2024).

The application of GPT-4 to financial statement analysis and earnings prediction in China's A-share market yields intriguing results that merit extensive discussion (Kim et al., 2024). The model's fluctuating performance, characterized by variable prediction accuracy and inconsistent correlation with stock returns, underscores the complexities inherent in applying large language models (LLMs) to financial markets, particularly in emerging economies (Wang and Di Iorio, 2007). The persistent high confidence levels exhibited by GPT-4, juxtaposed against its inconsistent accuracy, raise critical questions about the calibration of AI models in financial contexts and their potential for overconfidence bias. This phenomenon echoes broader concerns in the field of AI-driven financial analysis, where the opacity of model decision-making processes can lead to unwarranted trust in algorithmic predictions (Xing, 2024). The study's findings resonate with ongoing debates in behavioral finance regarding the limits of rational market theories, suggesting that even sophisticated AI models may struggle to consistently outperform in markets driven by policy interventions, retail investor sentiment, and unique structural characteristics (Kim et al., 2024).

The variable performance of GPT-4 across different market conditions highlights the challenge of developing robust AI models capable of adapting to the dynamic nature of financial markets (Wang and Di Iorio, 2007). This adaptability issue is particularly salient in the context of China's A-share market, where rapid economic transitions, regulatory shifts, and global economic integrations create a constantly evolving landscape. The model's performance fluctuations may be indicative of its sensitivity to changes in market microstructure, liquidity conditions, and macroeconomic factors, elements that are crucial in the broader discourse on market efficiency and the effectiveness of quantitative investment strategies (Kim et al., 2024). Moreover, the study's results contribute to the ongoing debate on the value of fundamental analysis in an era of high-frequency trading and alternative data sources. The inconsistent relationship between GPT-4's earnings predictions and subsequent stock returns challenges traditional assumptions about the linear relationship between fundamental analysis and market performance, echoing recent research on the diminishing returns of public financial information in increasingly efficient markets (Kim et al., 2024).

The ethical and regulatory implications of deploying AI models like GPT-4 in financial markets are profound and multifaceted. The potential for these models to exacerbate market inefficiencies, create information asymmetries, or be exploited for market manipulation necessitates a reevaluation of current regulatory frameworks (Xing, 2024). This study's findings underscore the need for transparent and interpretable AI systems in finance, aligning with global initiatives for responsible AI development (Du et al., 2024). Furthermore, the application of LLMs to financial analysis intersects with broader discussions on the democratization of financial advice and the potential for AI to bridge information gaps in emerging markets. However, the model's inconsistent performance cautions against over-reliance on AI-generated financial insights, particularly in markets with unique characteristics like China's A-share market (Cao et al., 2024b).

The study's exploration of GPT-4's performance in predicting earnings changes and stock returns also contributes to the burgeoning field of AI-driven ESG (Environmental, Social, and Governance) analysis (Du et al., 2024). As sustainable investing gains prominence globally, the potential for LLMs to process vast amounts of unstructured ESG data and extract meaningful insights becomes increasingly relevant (Kim et al., 2024). The model's ability to analyze financial statements could potentially be extended to non-financial disclosures, offering a new tool for assessing corporate sustainability and governance practices. This application aligns with the growing interest in integrating ESG factors into fundamental analysis and the challenges of quantifying qualitative ESG information (Kim et al., 2024).

Limitations of this study primarily stem from its focus on a single market and a specific time period. The unique characteristics of China's A-share market, including its high proportion of retail investors, frequent policy interventions, and the prevalence of state-owned enterprises, may limit the generalizability of findings to other markets (Xing, 2024). Additionally, the study's reliance on historical data raises questions about the model's ability to adapt to unprecedented market events or structural changes (Kim et al., 2024). The black-box nature of GPT-4's decision-making process also presents challenges in fully understanding the factors driving its predictions, potentially limiting its practical applicability in regulated financial environments where model interpretability is crucial (Wang and Di Iorio, 2007).

Future research directions should explore the integration of real-time market data and alternative data sources to enhance the model's adaptability and predictive power. Comparative studies across different markets and asset classes could provide insights into the generalizability of LLM-based financial analysis. Investigation into hybrid models that combine the strengths of LLMs with traditional financial models and human expertise could yield more robust and reliable prediction systems. Furthermore, research into techniques for improving model calibration and reducing overconfidence bias in AI financial analysis tools is essential for their responsible deployment in real-world investment scenarios (Kim et al., 2024).

5. Conclusion

The inconsistent relationship between prediction accuracy and stock returns challenges conventional assumptions about market efficiency and the value of fundamental analysis in the A-share market context. This disconnect emphasizes the need for a more holistic approach to financial modeling that incorporates market sentiment, policy impacts, and behavioral factors alongside fundamental analysis. Our study contributes to the growing body of literature on AI applications in finance, particularly in emerging markets. It underscores the potential of large language models like GPT-4 to augment traditional financial analysis methods while also highlighting the limitations and risks associated with overreliance on AI-generated insights. Looking forward, these findings call for further research into hybrid models that combine AI capabilities with human expertise and traditional financial models. There is a pressing need for improved techniques in model calibration, interpretability, and adaptability to changing market conditions. Additionally, the ethical and regulatory implications of deploying AI in financial markets demand careful consideration and the development of robust governance frameworks (Cao et al., 2024b). In conclusion, while GPT-4 shows promise in analyzing China's A-share market, its variable performance underscores both the potential and limitations of AI in complex financial ecosystems. As the field evolves, responsible integration of AI models into financial analysis processes may fundamentally transform investment

strategies and decision-making in global markets, necessitating ongoing research, ethical vigilance, and adaptive regulatory approaches.

References

[1] Bhat, R., & Jain, B. (2024, June). Stock Price Trend Prediction using Emotion Analysis of Financial Headlines with Distilled LLM Model. In Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments (pp. 67-73).

[2] Cao, Y., Chen, Z., Pei, Q., Dimino, F., Ausiello, L., Kumar, P., ... & Ndiaye, P. M. (2024). RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data. arXiv preprint arXiv:2404.07452.

[3] Cao, Y., Chen, Z., Pei, Q., Kumar, P., Subbalakshmi, K. P., & Ndiaye, P. M. (2024). ECC Analyzer: Extract Trading Signal from Earnings Conference Calls using Large Language Model for Stock Performance Prediction. arXiv preprint arXiv:2404.18470.

[4] Chen, S., Green, T. C., Gulen, H., & Zhou, D. (2024). What Does ChatGPT Make of Historical Stock Returns? Extrapolation and Miscalibration in LLM Stock Return Forecasts. arXiv preprint arXiv:2409.11540.

[5] Chen, Y., Kelly, B. T., & Xiu, D. (2022). Expected returns and large language models. Available at SSRN 4416687.

[6] Chiu, I., & Hung, M. W. Finance-Specific Large Language Models: Advancing Sentiment Analysis and Return Prediction with Llama 2. Mao-Wei, Finance-Specific Large Language Models: Advancing Sentiment Analysis and Return Prediction with Llama, 2.

[7] Cook, T. R., Kazinnik, S., Hansen, A. L., & McAdam, P. (2023). Evaluating local language models: An application to financial earnings calls. Available at SSRN 4627143.

[8] Dolphin, R., Dursun, J., Chow, J., Blankenship, J., Adams, K., & Pike, Q. (2024). Extracting Structured Insights from Financial News: An Augmented LLM Driven Approach. arXiv preprint arXiv:2407.15788.

[9] Drinkall, F., Pierrehumbert, J. B., & Zohren, S. (2024). Traditional Methods Outperform Generative LLMs at Forecasting Credit Ratings. arXiv preprint arXiv:2407.17624.

[10] Du, T., Kanodia, A., Brunborg, H., Vafa, K., & Athey, S. (2024). LABOR-LLM: Language-Based Occupational Representations with Large Language Models. arXiv preprint arXiv:2406.17972.

[11] Jansen, M., Swinkels, L., & Zhou, W. (2021). Anomalies in the China A-share market. Pacific-Basin Finance Journal, 68, 101607.

[12] Kim, A., Muhn, M., & Nikolaev, V. (2024). Financial statement analysis with large language models. arXiv preprint arXiv:2407.17866.

[13] Kim, S., Kim, S., Kim, Y., Park, J., Kim, S., Kim, M., ... & Lee, Y. (2023, November). LLMs Analyzing the Analysts: Do BERT and GPT Extract More Value from Financial Analyst Reports?. In Proceedings of the Fourth ACM International Conference on AI in Finance (pp. 383-391).

[14] Kou, Z., Yu, H., Peng, J., & Chen, L. (2024). Automate Strategy Finding with LLM in Quant investment. arXiv preprint arXiv:2409.06289.

[15] Li, X., Li, Z., Shi, C., Xu, Y., Du, Q., Tan, M., ... & Lin, W. (2024). AlphaFin: Benchmarking Financial Analysis with Retrieval-Augmented Stock-Chain Framework. arXiv preprint arXiv:2403.12582.

[16] Ou, J. A., & Penman, S. H. (1989). Financial statement analysis and the prediction of stock returns. Journal of accounting and economics, 11(4), 295-329.

[17] Ouyang, K., Liu, Y., Li, S., Bao, R., Harimoto, K., & Sun, X. (2024). Modal-adaptive Knowledge-enhanced Graph-based Financial Prediction from Monetary Policy Conference Calls with LLM. arXiv preprint arXiv:2403.16055.

[18] Pan, X., Zhang, R., Li, K., Tang, L., & Zhang, H. (2019). A cross-sectional and comparative study on Chinese A-share and H-share stock markets. International Research Journal of Applied Finance, 10(2), 84-108.

[19] Srivastava, V. (2024). BAI-Arg LLM at the FinLLM Challenge Task: Earn While You Argue-Financial Argument Identification. In Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning (pp. 165-173).

[20] Wang, Y., & Di Iorio, A. (2007). The cross section of expected stock returns in the Chinese A-share market. Global finance journal, 17(3), 335-349.

[21] Xing, F. (2024). Designing heterogeneous llm agents for financial sentiment analysis. ACM Transactions on Management Information Systems.

[22] Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., ... & Liu, T. (2024). Revolutionizing finance with llms: An overview of applications and insights. arXiv preprint arXiv:2401.11641.

[23] Zhou, Y. (2024). Predicting CDS Spreads and Stock Returns with Weather Risk: A Study Utilizing Nlp/Llm and Ai Measures. Llm and Ai Measures.