GENERATIVE ARTIFICIAL INTELLIGENCE AND THE FUTURE OF FINANCIAL FORECASTING: EVIDENCE FROM LARGE LANGUAGE MODELS

Mfon Akpan *

* College of Business and Technology, Northeastern State University, Tahlequah, USA
Contact details: College of Business and Technology, Northeastern State University, 600 N Grand Ave, Tahlequah, OK 74464, USA

How to cite this paper: Akpan, M. (2025). Generative artificial intelligence and the future of financial forecasting: Evidence from large language models. *Risk Governance and Control: Financial Markets & Institutions, 15*(4), 142–153. https://doi.org/10.22495/rgcv15i4p13

ISSN Online: 2077-4303
ISSN Print: 2077-429X

Received: 18.06.2025
Revised: 25.09.2025; 28.11.2025
Accepted: 05.12.2025

JEL Classification: G17, C45, C53, M41
DOI: 10.22495/rgcv15i4p13

Abstract

The predictive abilities of generative artificial intelligence (AI) are changing the landscape for analytic workflows across sectors. Nevertheless, its capacity and implications for use cases in high-stakes, non-stationary environments — like financial markets — have been empirically under-researched (Tan et al., 2023; Lin & Marques, 2024). This research paper examines generative AI's zero-shot forecasting capabilities using two large language model (LLM) architectures, OpenAI's GPT-4o, and Anthropic's Claude 3.5 Sonnet, as they forecast stock prices. Specifically, the paper evaluates the LLMs' predictive powers in terms of actual closing prices for a portfolio of in-use equities across sectors on February 3, 2025. A rigorous quantitative approach is used throughout the analyses. In the results section, standardized metrics including mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), correlation scores, and R-squared are calculated to assess predictive accuracy and directional bias. Results show that Claude 3.5 Sonnet outperformed GPT-4o on all accuracy metrics, and also showed better accuracy in forecasting actual movement in the stock market, confirming the study hypothesis and demonstrating performance can vary significantly between LLM architectures (Xu et al., 2024). Further analysis of sector-based performance can be undertaken. The study concludes that while the LLM Claude 3.5 Sonnet does yield encouraging strategic implications for use in investment analytics, there still exist significant challenges in relation to interpretability, model calibration, and model sensitivity to rapidly changing market dynamics.

Keywords: Generative Artificial Intelligence, Financial Forecasting, Stock Price Prediction, Large Language Models (LLMs), Zero-Shot Learning

Authors' individual contribution: The Author is responsible for all the contributions to the paper according to CRediT (Contributor Roles Taxonomy) standards.

Declaration of conflicting interests: The Author declares that there is no conflict of interest.

1. INTRODUCTION

The use of artificial intelligence (AI) in capital markets is a revolutionary evolution in the use of financial data for financial analytics, specifically predictive modeling. Current large language models (LLMs), such as OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet, use natural language processing and deep-learning architectures to aggregate financial data and provide predictive insights (Buczyński et al., 2023). These generative AI systems evaluate large quantities of historical stock data, economic signals, and sentiment data, producing zero-shot forecasts without human specialization or fine-tuning. Yet, note that Lin and Marques (2024) argue, the empirical usefulness of such models against stock prices is under-researched, especially with regard to some key components of predictive modeling: predictive accuracy, systematic bias, and statistical reliability.

Forecasting in the financial markets represents an ongoing analytical challenge for two main reasons: the stochastic, non-stationary nature of asset prices, combined with the multiplicity of factors coming from both fundamental economic factors and behavioral attributes. For some time, practitioners have used traditional forecasting methodologies to inform their decision-making, including both fundamental and technical analysis, as part of their investment process (Billah et al., 2024). As of late, the introduction of AI-powered model forecasts represents a different style of forecasting that assists decision makers with fully autonomous, high-frequency forecasting that may configure higher performance or potentially outperform existing methodologies. Previous research has started to investigate applications of AI to finance; however, the literature continues to lack actual empirical investigation on direct forecasting comparisons of the state-of-the-art LLMs like GPT-4o and Claude 3.5 Sonnet in a zero-shot forecasting situation. To date, the majority of research has either preferred to tune models in terms of older LLMs or simply applied LLMs similar to a fine-tune style model. Very little literature in either case has examined the capabilities of new, multi-purpose, and generally available LLMs.

This qualitative nature of this research aims to bridge the gap by holistically evaluating and assessing the predictive accuracy of GPT-4o and Claude 3.5 Sonnet based on live market data from diverse sectors. The primary aim of the research is to ascertain the validity of the predictive performance of these LLMs from the perspective of finance analysts and investors. Three main questions will guide the research and investigation:

RQ1: Which large language model, GPT-4o or Claude 3.5 Sonnet, is better able to accurately predict in predicting stock price performance under a zero-shot forecasting approach?

RQ2: What is the directional bias (underestimation or overestimation) in the predictions made by each model?

RQ3: How does the models' predictive accuracy differ?

This research uses a range of robust statistical methods, namely, mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), bias diagnostic statistics, correlation coefficients, and regression analysis to assess the model performance (El-Azab et al., 2024). These measurements give a balanced view of accuracy, variation, bias direction, and explanation. An evaluation of the deterioration of bias gives insight into how equity values are systematically misestimated over- or under. This is important for institutional analysts and retail investors.

By evaluating the predictive validity of LLMs in estimating stock prices, this study adds to the expanding academic discussion on the uses of AI in financial markets. The results will affect theoretical modeling and applied finance, and research in particular implications for portfolio strategy, risk management, and algorithmic trading. This study also informs professionals in corporate finance and data scientists about the possibilities and limits of applying LLMs for market estimation in an AI-dominated financial ecosystem.

The organization of the paper is as follows. Section 2 presents a detailed literature overview on AI in finance, traditional forecasting techniques, and the nascent field of LLM prediction. Section 3 explains the research methodology, including the data selection, prompting strategy for the LLMs, and evaluation techniques. Section 4 presents the empirical findings of the analysis. Section 5 discusses the findings, interpreting the results in relation to the literature and debating their implications. Finally, Section 6 concludes the paper, summarizing the main takeaways, suggesting some criticisms, and future research.

2. LITERATURE REVIEW

The intersection of AI and financial forecasting has garnered increasing scholarly attention, reflecting a convergence of disciplines, including data science, behavioral finance, and econometrics. Recent literature has documented a paradigm shift in forecasting methodologies, with AI-driven systems and LLMs emerging as viable alternatives to conventional statistical models (Cao, 2020). These models utilize natural language processing and machine learning algorithms to parse, interpret, and synthesize structured and unstructured financial data, enhancing predictive capabilities in volatile markets (Heaton et al., 2017; Das & Rad, 2020; Tan et al., 2023). In recent years, there have been examples of LLMs being used to forecast related stock price movement from financial news data, significantly outperforming the steep decline and relative stagnation of relevant traditional benchmark models, which demonstrates the disruptive potential of LLMs.

2.1. Artificial intelligence models in financial forecasting

Traditional forecasting techniques, such as autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroskedasticity (GARCH), have historically provided valuable insight into time series analysis (Bennell & Sutcliffe, 2011). However, their reliance on stationarity assumptions and linearity constraints has limited their applicability in increasingly complex and nonlinear financial environments (Buczyński et al., 2023). In contrast, as discussed by Nelson et al. (2017) and Duane et al. (2025), AI models can incorporate diverse data streams, including real-time sentiment data, macroeconomic indicators, and corporate disclosures, to produce more dynamic forecasts.

Recent developments in deep learning, notably recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, have demonstrated promising results in forecasting tasks (Xu et al., 2024). These models capture temporal dependencies and sequential patterns often lost in traditional frameworks (Hiransha et al., 2018). Moreover, LLMs such as GPT-4o and Claude 3.5 Sonnet extend this capability by integrating natural language processing with financial reasoning (Fischer & Krauss, 2018; Tang et al., 2025). Hence, enabling semantic interpretation of news reports, earnings calls, and regulatory filings.

Nonetheless, extant literature underscores inherent challenges in AI forecasting, including the risks of overfitting, model interpretability, and bias introduced through training data (Krollner et al., 2010). Studies such as Cao (2020) and Asamoah and Ahmed (2024) have also noted discrepancies between AI-generated predictions and actual market prices attributed to non-stationarity in time series data and model-specific idiosyncrasies.

2.2. Model evaluation and forecast accuracy

Evaluating AI models' predictive performance necessitates applying standardized statistical metrics. The most widely adopted include MAE, RMSE, and MAPE, which collectively assess model precision, dispersion, and proportionality of forecast deviations (Hu et al., 2021). The mean absolute scaled error (MASE) has been developed for comparing forecasts across a set of forecasts across forecasts with different time series (Sako et al., 2022), but there is still limited research using MASE in measuring AI-driven related financial predictions. Comparative studies have illustrated variability in model performance based on market conditions, data granularity, and temporal windows (Bennell & Sutcliffe, 2011; Chen et al., 2023). These metrics are essential in determining the efficacy of AI-generated forecasts and their economic value when integrated into investment decision-making frameworks.

Moreover, integrating hybrid models, combining AI-driven insights with technical and fundamental analysis, has been proposed to enhance prediction accuracy and mitigate risk exposure (Chen et al., 2023; Taheri Hosseinkhani, 2025). Such models offer a more holistic understanding of market dynamics and foster resilience against exogenous shocks.

2.3. Artificial intelligence bias and model transparency

The presence of bias in AI-generated forecasts presents a critical concern for academic researchers and financial practitioners. Bias may manifest as systematic overestimation or underestimation, often rooted in historical data imbalances or algorithmic design (Heaton et al., 2017). According to Rajanand and Singh (2023), different LLM architectures prioritize distinct information streams. For example, some favor sentiment analysis while others emphasize quantitative financial ratios. The implications of such disparities are non-trivial, especially in high-frequency trading environments where milliseconds of miscalculation can result in substantial financial loss. For example, this architectural bias can produce fundamentally different performance in financial contexts across models with similar capacities, leading the literature to examine alternatives to bias.

Furthermore, explainability or the degree to which AI model decisions can be interpreted remains a frontier concern (Hu et al., 2021). Bahoo et al. (2024) state that as financial institutions increasingly adopt AI for predictive modeling, regulatory frameworks may necessitate greater model architecture and decision logic transparency to ensure compliance and maintain investor trust. Furthermore, the broader notion of explainability and trust, which is defined as the clear scientific reasoning when applied to complex LLM-type models, remains a significant barrier to adoption for use in regulated financial environments, driving research around explainable AI solutions for finance.

2.4. Synthesis and research gap

The reviewed literature substantiates the growing influence of AI models in financial forecasting and provides insight into their outcomes' complexity and variability. While AI systems, including LLMs, offer enhanced analytical capacity, their performance is contingent upon input quality, model calibration, and contextual alignment. There is a recognized gap in empirical studies examining the zero-shot predictive performance of advanced LLMs such as GPT-4o and Claude 3.5 Sonnet across a multi-sectoral portfolio. Most of the literature has explored related capabilities in fine-tuned models or in earlier generations or different classes of neural network architectures (convolutional neural networks, RNNs), leaving the capabilities of newer generalized versions of LLMs as yet unexamined. Additionally, the literature is sparse on comparative analyses that systematically evaluate LLMs' bias, correlation, and predictive strength under real-world conditions. Accordingly, this study endeavors to fill that gap by applying a rigorous statistical framework to benchmark and assess the predictive validity of these models against actual market data. The research contributes to a more nuanced understanding of AI's role and limitations in contemporary financial modelling.

3. METHODOLOGY

3.1. Research design

This study employs a quantitative, comparative research design to evaluate the predictive accuracy of generative AI models in financial market forecasting. The primary objective is to assess and compare the zero-shot forecasting capabilities of two LLMs, namely OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet, against actual stock market data. A zero-shot framework is instrumental in testing a model's ability to generate predictions without domain-specific fine-tuning or prior contextual training. This approach enhances the objectivity of the evaluation and reflects a realistic deployment scenario in which financial analysts may utilize publicly available models without proprietary enhancements.

The design is structured to address core research questions related to forecast accuracy, directional bias, correlation with actual market prices, and model robustness across multiple industries. Its comparative nature ensures that differences in performance are systematically measured using a consistent and replicable methodological framework.

3.2. Data collection procedures

Historical stock price data were sourced directly from Google Finance and extracted through Google Sheets using an automated Google Apps Script. The following formula was used for data retrieval:

```
= GOOGLEFINANCE ("Ticker Symbol", "close",
"1/1/2024", "12/31/2024", "DAILY")
```

This method provided a reliable, reproducible dataset covering each selected company's 2024 calendar year. By leveraging publicly accessible tools, the data collection process maintains transparency and scalability while aligning with best practices in financial data acquisition.

Stock price forecasts were generated using identical prompts submitted to GPT-4o and Claude 3.5 Sonnet. Both models were queried under zero-shot conditions without access to real-time market data or external plugins. Prompts were standardized to elicit stock price predictions for specific companies within a defined timeframe. This prompt-driven mechanism ensured uniformity and eliminated bias introduced by user-guided model interactions.

3.3. Industry and firm selection

Companies were selected across ten major industry sectors to ensure broad market representation and enhance generalizability, reflecting varying levels of market volatility, regulatory complexity, and capital structure. Each sector includes five representative firms:

Table 1. Industry and firm selection

Industry	***Companies***
Technology	Apple Inc. (AAPL), Microsoft Corp. (MSFT), NVIDIA Corp. (NVDA), Intel Corp. (INTC), Adobe Inc. (ADBE), Alphabet Inc. (GOOG), Meta Platforms Inc. (META)
Healthcare	Johnson & Johnson (JNJ), Pfizer Inc. (PFE), UnitedHealth Group (UNH), Abbott Laboratories (ABT), Moderna Inc. (MRNA)
Financial services	JPMorgan Chase & Co. (JPM), Bank of America (BAC), Goldman Sachs Group (GS), Wells Fargo & Co. (WFC), Citigroup Inc. (C)
Consumer discretionary	Amazon.com Inc. (AMZN), Tesla Inc. (TSLA), Home Depot Inc. (HD), McDonald's Corp. (MCD), Nike Inc. (NKE)
Consumer staples	Procter & Gamble Co. (PG), Coca-Cola Co. (KO), PepsiCo Inc. (PEP), Walmart Inc. (WMT), Costco Wholesale Corp. (COST)
Energy	ExxonMobil Corp. (XOM), Chevron Corp. (CVX), ConocoPhillips (COP), Schlumberger Ltd. (SLB), Marathon Petroleum Corp. (MPC)
Industrials	Boeing Co. (BA), Caterpillar Inc. (CAT), Union Pacific Corp. (UNP), Honeywell International Inc. (HON), General Electric Co. (GE)
Materials	Dow Inc. (DOW), Newmont Corp. (NEM), Freeport-McMoRan Inc. (FCX), Linde plc (LIN), Albemarle Corp. (ALB)
Utilities	Duke Energy Corp. (DUK), Dominion Energy Inc. (D), NextEra Energy Inc. (NEE), Southern Co. (SO), Consolidated Edison Inc. (ED)
Real estate	Simon Property Group Inc. (SPG), American Tower Corp. (AMT), Prologis Inc. (PLD)

Note: This diversified portfolio enables cross-sectoral analysis and reveals how AI model performance may vary based on underlying firm and industry characteristics.

3.4. Experimental procedure

Each AI model received a zero-shot prompt requesting the selected firms' end-of-year 2024 stock price predictions. No contextual training, macroeconomic background, or real-time market data was provided. The models' output prices were documented and compared to the actual closing prices obtained from Google Finance.

The performance of each model was then evaluated using the following statistical metrics:

- MAE: Captures the average magnitude of prediction errors.
- RMSE: Penalizes more significant deviations, emphasizing high-variance discrepancies.
- MAPE: This expression expresses forecast error as a percentage of actual values, normalizing across firms.
- Bias analysis: Assesses whether models consistently overestimate or underestimate stock prices.
- Pearson correlation coefficient (r): Measures the strength and direction of association between predicted and actual values.
- Linear regression analysis (R^2): Quantifies the extent to which model predictions explain the variance in actual stock prices.

3.5. Data analysis and statistical tools

All data were structured and preprocessed using Python, with analytical computations performed using Python's pandas, numpy, and scikit-learn libraries and supplemental visualization via matplotlib. Statistical significance testing and regression modeling were conducted using R and MATLAB for validation and robustness. Model performance was examined at both individual firm and aggregated industry levels.

The analysis followed this structured process:

1. Data cleaning and normalization involve removing missing values, standardizing date formats, and aligning prediction timestamps.
2. Error computation: Calculate each model-firm pair's MAE, RMSE, and MAPE.
3. Bias detection: Estimation of directional error tendencies through residual plots.
4. Correlation and regression modeling: Quantitative evaluation of model explanatory power and trend alignment.

3.6. Reliability, validity, and ethical considerations

All data sources and statistical methods were transparent and replicable to ensure methodological reliability. The study used a standardized prompt mechanism across both models to mitigate interaction-based variance. Multiple statistical metrics and confirming results across diverse industries strengthened validity.

AI tools, specifically OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet, were used to generate zero-shot stock price forecasts based on standardized prompts. Analytical computations were initially performed using an LLM (OpenAI's ChatGPT), which executed Python-based analysis leveraging pandas for data handling, NumPy for numerical operations, and scikit-learn for model performance metrics. The following statistical measures were calculated for each AI model: MAE, RMSE, MAPE, bias, Pearson correlation coefficient (r), and coefficient of determination (R^2). To ensure analytical rigor and methodological integrity, all results were independently verified and validated using RStudio.

No human subjects were involved, and all data were publicly sourced, so no institutional review board approval was required. Nonetheless, ethical considerations were adhered to in the transparent reporting of results, acknowledgment of limitations, and avoidance of model misrepresentation.

3.7. Consideration of alternative methods

While the selected quantitative, zero-shot comparative design is a suitable means for answering the research questions, there are other methodological perspectives that could be applied to research with a similar intention. A fine-tuning approach might be an execution of a form of the modeling process, meaning the base LLMs (GPT-4o and Claude 3.5 Sonnet) are trained further on a specific dataset relating to historical stock prices and financial news (Nasiopoulos et al., 2025). With fine-tuning of the model, accuracy might improve for the particular task; however, it would also be an effective confound, making it challenging to disentangle the base model's inherent predictive capability from the observed influences of new training data. Therefore, for the purposes of our structurally zero-shot design, we could intentionally steer clear of this confound in order to test the models' out-of-the-box reasoning.

As an alternative approach, a qualitative or mixed-methods design might be implemented. During this kind of research, experts could be interviewed about their experiences with financial analysts to elicit explanations for the LLMs' reasoning processes or to situate the errors made in a quantitative analysis. Although this would provide rich contextual explanation, it would not produce the objective, generalizable, and statistically testable outcomes associated with a quantitative comparison. Lastly, additional benchmark models can be used, given that there are several traditional time-series models that we could have used for a benchmark comparison with the LLMs' predictions, such as ARIMA or alternative AI models like LSTMs. This study aims to try a practice-based head-to-head comparison of the two most advanced general-purpose LLMs to provide practitioners with a pragmatic, contemporary benchmark for the potential use of these LLMs in their practice. The aforementioned models could easily be added as benchmarks for future research.

4. RESULTS

This section presents the empirical findings from the head-to-head comparison between GPT-4o and Claude 3.5 Sonnet, based on their zero-shot stock price forecasts for a diversified portfolio of over 50 firms. Each model's prediction was compared against the actual closing stock price on February 3, 2025, as retrieved from Google Finance. Performance was evaluated using six core statistical metrics: MAE, RMSE, MAPE, bias, Pearson correlation coefficient (r), and R-squared (R^2).

Table 2. Model performance metrics

Metric	*GPT-4o*	*Claude 3.5 Sonnet*
MAE	29.98	24.39
RMSE	93.17	82.82
MAPE	9.10	8.55
Bias	-23.82	-15.54
r	0.89	0.91
R^2	0.77	0.82

Note: All metrics include outlier data. Predictions for high-value, high-volatility firms — such as GS and COST — resulted in significant errors for both models. Their inclusion deliberately reflected the real-world complexity and volatility inherent in financial forecasting.

Therefore, the results may be summarized as follows:

- Claude 3.5 Sonnet demonstrated superior performance across most key metrics. It achieved lower MAE and RMSE, indicating fewer minor and significant prediction errors. This means Claude 3.5 Sonnet's forecasts were, on average, closer to the actual stock prices.
- Bias was also significantly lower for Claude 3.5 Sonnet (-15.54) than GPT-4o (-23.82), showing that Claude 3.5 Sonnet had a more neutral error pattern. GPT-4o exhibited a stronger tendency to underpredict, particularly for higher-priced equities.
- Both models' Pearson correlation (r) and R^2 values were high, indicating strong linear alignment between predicted and actual prices. Claude 3.5 Sonnet had a slightly higher correlation (r = 0.91) and better explanatory power ($R^2 = 0.82$), confirming its forecasts aligned more closely with market trends.
- Interestingly, GPT-4o delivered a lower MAPE, suggesting it performed relatively better when measuring percentage-based accuracy, particularly on lower-priced stocks. This makes GPT-4o a potential fit for use cases where proportional accuracy across price scales is prioritized.

5. DISCUSSION

This research assessed the zero-shot forecasting accuracy of two cutting-edge LLMs, GPT-4o and Claude 3.5 Sonnet, in predicting stock prices. The results suggest that both LLMs exhibited a high predictive power overall, but their levels of accuracy and bias were not equivalent, with Claude 3.5 Sonnet overall being more accurate with lower bias. This study's findings have substantial implications for the theoretical knowledge of LLMs in finance as well as their application and practice.

5.1. Discussion of key findings

The superior performance demonstrated by Claude 3.5 Sonnet, reflected in its lower MAE and RMSE, is indicative of a more powerful internal modeling of financial relationships. This is in agreement with the body of literature that suggests more sophisticated AI architectures are able to better understand the non-linear and complex relationships present in financial markets (Buczyński et al., 2023; Xu et al., 2024). Furthermore, it is especially interesting that this was achieved in a zero-shot setting. The work extends that of Tan et al. (2023), in which they offered evidence of LLMs' predictive ability from financial news, so we are now able to show that these abilities extend even further into forecast tasks driven by numerics, without the need for task-specific fine-tuning.

A systematic negative bias, or underprediction, is an important finding in both models, although it is slightly more pronounced in GPT-4o. This observation is consistent with the known problems of AI bias, which are tied to how algorithms are designed and trained, and the information that is included in those data (Heaton et al., 2017). The underprediction could be characterized as a conservative, risk-averse approach that was there for certain outcomes due to the amounts of commentary, or market crash and decline-related material, that was likely included as an element of the models that were trained using the language in the model's training data. This relates to Rajanand and Singh's (2023) comments that LLM architectures can, and do, prioritize the information streams differently than others. Our results would imply that Claude 3.5 Sonnet, as an architecture, appears relatively less biased in some specific ways.

5.2. Discussion in the context of existing literature

The elevated correlation coefficients and R^2 values for both models support the ability of LLMs to discover and emulate underlying market trends, a long-standing aim of quantitative finance (Fischer & Krauss, 2018). Conversely, our findings further substantiate Krollner et al.'s (2010) and Cao's (2020) empirical claims about model-specific idiosyncrasies and model-specific performance. The difference in performance between two models of similar quality indicates that "AI" is not a homogeneous thing; model selection is an important driver of forecasting success.

The deliberate inclusion and inspection of outliers demonstrates an application rooted in real-world testing and a strategy hailed in the model evaluation literature (Hu et al., 2021). The relative robustness of Claude 3.5 Sonnet with high-volatility stocks like GS and COST can be viewed as a potential edge over traditional models like ARIMA that fail to respond to the "noise" and non-stationarity of the market (Bennell & Sutcliffe, 2011). This argues for a future of implementation of hybrid approaches leveraging the trend-capacity of LLMs and smoothing techniques from traditional econometrics, supported by recent work from Chen et al. (2023) and Taheri Hosseinkhani (2025).

5.3. Practical implications

These results are particularly useful for financial analysts and portfolio managers from a practical perspective. The potential for zero-shot forecasting suggests that these new tools can be accessed with nearly no technical barrier/cost for fine-tuning. Claude 3.5 Sonnet can complement analytical workflows for matters such as generating a first-order forecast, a stress-tested scenario, or some underlying market anomaly, all with greater accuracy and less bias. Despite that, there continue to be error-related challenges, especially in cases with outliers, and the black box nature of model predictions can be problematic for future usage and updates. As Bahoo et al. (2024) and Hu et al. (2021) highlighted, transparency is particularly important for regulatory compliance and trust. These models should still be useful, but for the near term, they must be used as a tool to augment rather than to replace human judgment.

6. CONCLUSION

This study compared the zero-shot forecasting performance of GPT-4o and Claude 3.5 Sonnet across a cross-industry portfolio of publicly traded companies. Each model was given a standardized prompt and tasked with forecasting the closing stock price as of February 3, 2025, using historical pricing data from 2024. The findings demonstrate that Claude 3.5 Sonnet outperformed GPT-4o in absolute accuracy (MAE and RMSE), directional stability (bias), and linear alignment (R^2 and r). Claude's predictions were closer to actual values and better explained observed market trends. GPT-4o, while competitive in relative percentage accuracy, showed higher variance and a pronounced underprediction tendency, particularly on higher-value stocks. Both models strongly correlate with actual prices, suggesting that LLMs can capture macro-level market patterns without access to real-time financial feeds or fine-tuning. These results reinforce the growing potential of generative AI to augment traditional forecasting workflows, mainly when used alongside human analysts or structured models.

The theoretical implications of this study both contribute to the significant literature on LLM applicability in finance and reaffirm the substantial disruptive power of generative AI in the realm of financial analytic solutions. Importantly, striking differences in performance were found between different model architectures, which speaks to the significance of model selection as an important and non-trivial factor in the profession. Limitations, by nature of design, are acknowledged, which also create opportunities for future inquiries. In particular, the analysis focused on a singular forecasting horizon, and while the rigid zero-shot framework is valuable to assess capabilities, it does not utilize domain-fine-tuning nor direct ingestion of data that may contribute to more precise forecasts. Additionally, models were analyzed without real-time incorporation of news sentiment or macroeconomic indicators, known to influence market movement.

Based on the implications and limitations of the findings, a few possible directions for future research can be identified. Future studies should assess multi-step role forecasting with longer quarterly or yearly horizons, research the effects of sector-specific fine-tuning, and research ensemble methods that combine LLM outputs with more traditional time-series models. Research should also integrate macroeconomic indicators, sentiment analysis using news feeds, and test in additional international markets and additional asset classes to address the question of the generalizability of the findings. From a practitioner perspective, the findings indicate the preferred use of Claude 3.5 Sonnet within decision-support systems for portfolio modeling and risk analytics. In the process of acquiring these capabilities, financial firms should consider developing sector-specific prompt calibration protocols to achieve a more granular level of precision in financial return potential, implement robust outlier detection systems to identify flags in outlier predictions for human review, and implement ongoing model governance practices to monitor for performance drift as markets evolve. The implication of this research is that the best application of these powerful generative AI models is as augmenting engines instead of oracle tools; these models can augment the speed, breadth, and depth of human financial experience, but their use should be framed through processes that facilitate transparency and explainability to support compliance and trust in pre-scribed financial environments.

Future research can pursue several directions: first, investigating the rationale of the predictions by utilizing prompting techniques with a chain-of-thinking to increase explainability. Second, examining the effects of incorporating a news feed (realistic) or fundamental information as a part of the prompt. Finally, longer-term work also requires further research to understand how the model changes during bullish or bearish market regimes.

REFERENCES

Asamoah, P. B., & Ahmed, I. A. (2024). AI-assisted forecasting of energy prices. *International Journal of Research and Scientific Innovation, 11*(9), 626–649. https://doi.org/10.51244/IJRSI.2024.1109057

Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in finance: A comprehensive review through bibliometric and content analysis. *SN Business & Economics*, *4,* Article 23. https://doi.org/10.1007/s43546-023-00618-x

Bennell, J. A., & Sutcliffe, C. (2011). *Black-scholes versus artificial neural networks in pricing FTSE 100 options.* https://doi.org/10.2139/ssrn.544882

Billah, M. M., Sultana, A., Bhuiyan, F., & Kaosar, M. G. (2024). Stock price prediction: Comparison of different moving average techniques using deep learning model. *Neural Computing and Applications*, *36,* 5861–5871. https://doi.org/10.1007/s00521-023-09369-0

Buczyński, M., Chlebus, M., Kopczewska, K., & Zajenkowski, M. (2023). Financial time series models — Comprehensive review of deep learning approaches and practical recommendations. *Engineering Proceedings*, *39*(1), Article 79. https://doi.org/10.3390/engproc2023039079

Cao, L. (2020). *AI in finance: A review.* https://doi.org/10.2139/ssrn.3647625

Chen, W., Hussain, W., Cauteruccio, F., & Zhang, X. (2023). Deep learning for financial time series prediction: A state-of-the-art review of standalone and hybrid models. *Computer Modeling in Engineering & Sciences, 139*(1), 187–224. https://doi.org/10.32604/cmes.2023.031388

Chen, Z., Balan, M. M., & Brown, K. (2023). *Language models are few-shot learners for prognostic prediction.* ArXiv. https://doi.org/10.32604/cmes.2023.031388

Das, A., & Rad, P. (2020). *Opportunities and challenges in explainable artificial intelligence (XAI): A survey*. ArXiv. https://arxiv.org/abs/2006.11371

Duane, J., Morgan, A., & Carter, E. (2025). *A review of financial data analysis techniques for unstructured data in the deep learning era: Methods, challenges, and applications*. OSF Preprints. https://doi.org/10.31219/osf.io/gdvbj_v1

El-Azab, H.-A. I., Swief, R. A., El-Amary, N. H., & Temraz, H. K. (2024). Machine and deep learning approaches for forecasting electricity price and energy load assessment on real datasets. *Ain Shams Engineering Journal*, *15*(4), Article 102613. https://doi.org/10.1016/j.asej.2023.102613

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci, 6*(1), Article 3. https://doi.org/10.3390/sci6010003

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Ghosh, B. P., Bhuiyan, M. S., Das, D., Nguyen, T. N., Jewel, R. M., Mia, M. T., Cao, D. M., & Shahid, R. (2024). Deep learning in stock market forecasting: Comparative analysis of neural network architectures across NSE and NYSE. *Journal of Computer Science and Technology Studies*, *6*(1), 68–75. https://doi.org/10.32996/jcsts.2024.6.1.8

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry, 33*(1), 3–12. https://doi.org/10.1002/asmb.2209

Hiransha, M., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2018). NSE stock market prediction using deep-learning models. *Procedia Computer Science, 132*, 1351–1362. https://doi.org/10.1016/j.procs.2018.05.050

Hu, Z., Zhao, Y., & Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation, 4*(1), Article 9. https://doi.org/10.3390/ASI4010009

Jasni, N. S., & Zulkifli, A. (2024). The moderating role of sector risk in the relationship between ESG and financial performance: Evidence from top companies in Malaysia. *Edelweiss Applied Science and Technology*, *8*(2), 59–72. https://doi.org/10.55214/25768484.v8i2.672

Krollner, B., Vanstone, B., & Finnie, G. (2010). Financial time series forecasting with machine learning techniques: A survey. In *Proceedings of the 18th European Symposium on Artificial Neural Networks: Computational Intelligence and Machine Learning* (pp. 25–30). Bond University. https://surl.li/qgndfo

Lin, C. Y., & Marques, J. A. L. (2024). Stock market prediction using artificial intelligence: A systematic review of systematic reviews. *Social Sciences & Humanities Open, 9* Article 100864. https://doi.org/10.1016/j.ssaho.2024.100864

Nasiopoulos, D. K., Roumeliotis, K. I., Sakas, D. P., Toudas, K., & Reklitis, P. (2025). Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models. *International Journal of Financial Studies*, *13*(2), Article 75. https://doi.org/10.3390/ijfs13020075

Nelson, D. M. Q., Pereira, A. C. M., & de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. In *Proceedings of the 2017 International Joint Conference on Neural Networks* (pp. 1419–1426). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/IJCNN.2017.7966019

Rajanand, A., & Singh, P. (2023). Stock price prediction using depthwise pointwise CNN with sequential LSTM. In *Proceedings of the 2023 2nd International Conference on Applied Artificial Intelligence and Computing* (pp. 82–86). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/ICAAIC56838.2023.10140728

Sako, K., Mpinda, B. N., & Rodrigues, P. C. (2022). Neural networks for financial time series forecasting. *Entropy*, *24*(5), Article 657. https://doi.org/10.3390/e24050657

Sangeetha, J. M., & Alfia, K. J. (2024). Financial stock market forecast using evaluated linear regression based machine learning technique. *Measurement: Sensors*, *31*, Article 100950. https://doi.org/10.1016/j.measen.2023.100950

Taheri Hosseinkhani, N. (2025). *Artificial intelligence applications in financial markets and corporate finance: Technologies, challenges, and opportunities*. https://doi.org/10.2139/ssrn.5403522

Tan, L., Wu, H., & Zhang, X. (2023). *Large language models and return prediction in China*. https://doi.org/10.2139/ssrn.4712248

Tang, Z., Haihong, E., Ma, Z., He, H., Liu, J., Yang, Z., Rong, Z., Li, R., Ji, K., Huang, Q., Hu, X., Liu, Y., & Zheng, Q. (2025). Finance reasoning: Benchmarking financial numerical reasoning more credible, comprehensive and challenging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 15721–15749). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.766

Taveekitworachai, P., Abdullah, F., & Thawonmas, R. (2024). *Large language models are null-shot learners*. arXiv. https://arxiv.org/abs/2401.08273

Thrun, M. C. (2022). Exploiting distance-based structures in data using an explainable AI for stock picking. *Information*, *13*(2), Article 51. https://doi.org/10.3390/info13020051

Vidya Sagar, P., Rajyalaxmi, M., Subbalakshmi, A. V. V. S., Sengan, S., Bommisetti, R. K., & Dadheech, P. (2024). Utilizing stochastic differential equations and random forest for precision forecasting in stock market dynamics. *Journal of Interdisciplinary Mathematics*, *27*(2), 285–298. https://doi.org/10.47974/JIM-1822

Widiputra, H., & Juwono, E. (2024). Parallel multivariate deep learning models for time-series prediction: A comparative analysis in Asian stock markets. *IAES International Journal of Artificial Intelligence*, *13*(1), 475–486. https://doi.org/10.11591/ijai.v13.i1.pp475-486

Xu, W., Chen, J., & Xiao, J. (2024, November 10). *A hybrid price forecasting model for the stock trading market based on AI technique*. TechRxiv. https://doi.org/10.36227/techrxiv.173121439.97240664/v1

APPENDIX A. GOOGLE SHEETS FORMULAS FOR HISTORICAL STOCK PRICE RETRIEVAL (2024)

Use the following formulas in google sheets to retrieve daily closing prices for each stock in your dataset:

Apple Inc. (AAPL)

=GOOGLEFINANCE("AAPL", "close", "1/1/2024", "12/31/2024", "DAILY")

Microsoft Corp. (MSFT)

=GOOGLEFINANCE("MSFT", "close", "1/1/2024", "12/31/2024", "DAILY")

NVIDIA Corp. (NVDA)

=GOOGLEFINANCE("NVDA", "close", "1/1/2024", "12/31/2024", "DAILY")

Intel Corp. (INTC)

=GOOGLEFINANCE("INTC", "close", "1/1/2024", "12/31/2024", "DAILY")

Adobe Inc. (ADBE)

=GOOGLEFINANCE("ADBE", "close", "1/1/2024", "12/31/2024", "DAILY")

Johnson & Johnson (JNJ)

=GOOGLEFINANCE("JNJ", "close", "1/1/2024", "12/31/2024", "DAILY")

Pfizer Inc. (PFE)

=GOOGLEFINANCE("PFE", "close", "1/1/2024", "12/31/2024", "DAILY")

UnitedHealth Group (UNH)

=GOOGLEFINANCE("UNH", "close", "1/1/2024", "12/31/2024", "DAILY")

Abbott Laboratories (ABT)

=GOOGLEFINANCE("ABT", "close", "1/1/2024", "12/31/2024", "DAILY")

Moderna Inc. (MRNA)

=GOOGLEFINANCE("MRNA", "close", "1/1/2024", "12/31/2024", "DAILY")

JPMorgan Chase & Co. (JPM)

=GOOGLEFINANCE("JPM", "close", "1/1/2024", "12/31/2024", "DAILY")

Bank of America (BAC)

=GOOGLEFINANCE("BAC", "close", "1/1/2024", "12/31/2024", "DAILY")

Goldman Sachs Group (GS)

=GOOGLEFINANCE("GS", "close", "1/1/2024", "12/31/2024", "DAILY")

Wells Fargo & Co. (WFC)

=GOOGLEFINANCE("WFC", "close", "1/1/2024", "12/31/2024", "DAILY")

Citigroup Inc. (C)

=GOOGLEFINANCE("C", "close", "1/1/2024", "12/31/2024", "DAILY")

Amazon.com Inc. (AMZN)

=GOOGLEFINANCE("AMZN", "close", "1/1/2024", "12/31/2024", "DAILY")

Tesla Inc. (TSLA)

=GOOGLEFINANCE("TSLA", "close", "1/1/2024", "12/31/2024", "DAILY")

Home Depot Inc. (HD)

=GOOGLEFINANCE("HD", "close", "1/1/2024", "12/31/2024", "DAILY")

McDonald’s Corp. (MCD)

=GOOGLEFINANCE("MCD", "close", "1/1/2024", "12/31/2024", "DAILY")

Nike Inc. (NKE)

=GOOGLEFINANCE("NKE", "close", "1/1/2024", "12/31/2024", "DAILY")

Procter & Gamble Co. (PG)

=GOOGLEFINANCE("PG", "close", "1/1/2024", "12/31/2024", "DAILY")

Coca-Cola Co. (KO)

=GOOGLEFINANCE("KO", "close", "1/1/2024", "12/31/2024", "DAILY")

PepsiCo Inc. (PEP)

=GOOGLEFINANCE("PEP", "close", "1/1/2024", "12/31/2024", "DAILY")

Walmart Inc. (WMT)

=GOOGLEFINANCE("WMT", "close", "1/1/2024", "12/31/2024", "DAILY")

Costco Wholesale Corp. (COST)

=GOOGLEFINANCE("COST", "close", "1/1/2024", "12/31/2024", "DAILY")

ExxonMobil Corp. (XOM)

=GOOGLEFINANCE("XOM", "close", "1/1/2024", "12/31/2024", "DAILY")

Chevron Corp. (CVX)

=GOOGLEFINANCE("CVX", "close", "1/1/2024", "12/31/2024", "DAILY")

ConocoPhillips (COP)

=GOOGLEFINANCE("COP", "close", "1/1/2024", "12/31/2024", "DAILY")

Schlumberger Ltd. (SLB)

=GOOGLEFINANCE("SLB", "close", "1/1/2024", "12/31/2024", "DAILY")

Marathon Petroleum Corp. (MPC)

=GOOGLEFINANCE("MPC", "close", "1/1/2024", "12/31/2024", "DAILY")

Boeing Co. (BA)

=GOOGLEFINANCE("BA", "close", "1/1/2024", "12/31/2024", "DAILY")

Caterpillar Inc. (CAT)

=GOOGLEFINANCE("CAT", "close", "1/1/2024", "12/31/2024", "DAILY")

Union Pacific Corp. (UNP)

=GOOGLEFINANCE("UNP", "close", "1/1/2024", "12/31/2024", "DAILY")

Honeywell International Inc. (HON)

=GOOGLEFINANCE("HON", "close", "1/1/2024", "12/31/2024", "DAILY")

General Electric Co. (GE)

=GOOGLEFINANCE("GE", "close", "1/1/2024", "12/31/2024", "DAILY")

Dow Inc. (DOW)

=GOOGLEFINANCE("DOW", "close", "1/1/2024", "12/31/2024", "DAILY")

Newmont Corp. (NEM)

=GOOGLEFINANCE("NEM", "close", "1/1/2024", "12/31/2024", "DAILY")

Freeport-McMoRan Inc. (FCX)

=GOOGLEFINANCE("FCX", "close", "1/1/2024", "12/31/2024", "DAILY")

Linde plc (LIN)

=GOOGLEFINANCE("LIN", "close", "1/1/2024", "12/31/2024", "DAILY")

Albemarle Corp. (ALB)

=GOOGLEFINANCE("ALB", "close", "1/1/2024", "12/31/2024", "DAILY")

Duke Energy Corp. (DUK)

=GOOGLEFINANCE("DUK", "close", "1/1/2024", "12/31/2024", "DAILY")

Dominion Energy Inc. (D)

=GOOGLEFINANCE("D", "close", "1/1/2024", "12/31/2024", "DAILY")

NextEra Energy Inc. (NEE)

=GOOGLEFINANCE("NEE", "close", "1/1/2024", "12/31/2024", "DAILY")

Southern Co. (SO)

=GOOGLEFINANCE("SO", "close", "1/1/2024", "12/31/2024", "DAILY")

Consolidated Edison Inc. (ED)

=GOOGLEFINANCE("ED", "close", "1/1/2024", "12/31/2024", "DAILY")

Simon Property Group Inc. (SPG)

=GOOGLEFINANCE("SPG", "close", "1/1/2024", "12/31/2024", "DAILY")

American Tower Corp. (AMT)

=GOOGLEFINANCE("AMT", "close", "1/1/2024", "12/31/2024", "DAILY")

Prologis Inc. (PLD)

=GOOGLEFINANCE("PLD", "close", "1/1/2024", "12/31/2024", "DAILY")

Alphabet Inc. (GOOG)

=GOOGLEFINANCE("GOOG", "close", "1/1/2024", "12/31/2024", "DAILY")

Meta Platforms Inc. (META)

=GOOGLEFINANCE("META", "close", "1/1/2024", "12/31/2024", "DAILY")

APPENDIX B. FULL STOCK PREDICTIONS AND ACTUAL PRICES

Table B.1. Stock predictions and actual closing prices

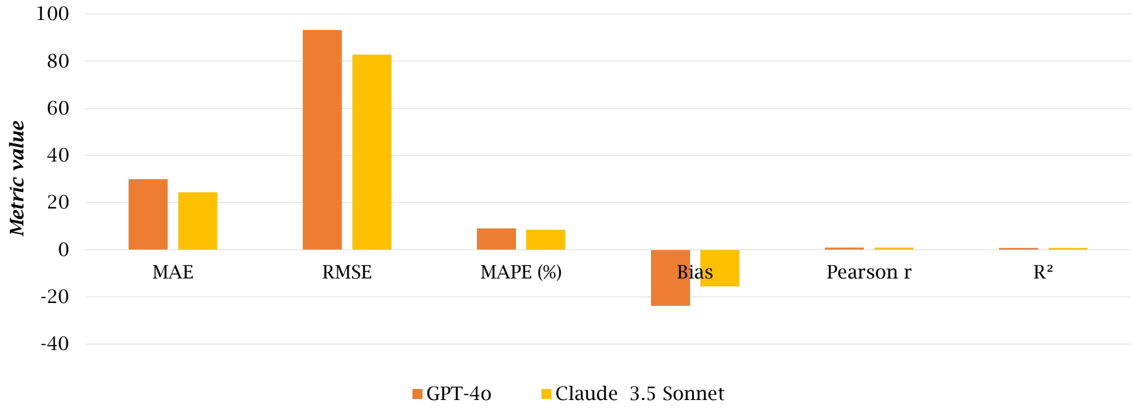
Ticker	*Company*	*GPT-4o prediction*	*Claude 3.5 Sonnet prediction*	*Actual closing price (February 3, 2025)*
AAPL	Apple Inc.	252.20	247.86	228.01
MSFT	Microsoft Corp.	416.66	428.75	410.92
NVDA	NVIDIA Corp.	137.46	151.78	116.66
INTC	Intel Corp.	20.02	21.15	19.38
ADBE	Adobe Inc.	445.19	495.82	438.60
JNJ	Johnson & Johnson	144.68	153.45	151.87
PFE	Pfizer Inc.	26.54	26.15	26.20
UNH	UnitedHealth Group	505.27	508.45	548.18
ABT	Abbott Laboratories	113.04	113.50	128.45
MRNA	Moderna Inc.	41.70	40.15	36.55
JPM	JPMorgan Chase & Co.	223.64	243.85	266.81
BAC	Bank of America	43.97	44.82	46.21
GS	Goldman Sachs Group	65.08	70.00	632.37
WFC	Wells Fargo & Co.	72.66	70.85	78.14
C	Citigroup Inc.	55.53	70.85	79.61
AMZN	Amazon.com Inc.	203.07	224.50	237.42
TSLA	Tesla Inc.	400.88	408.73	383.68
HD	Home Depot Inc.	391.90	396.45	408.76
MCD	McDonald’s Corp.	289.38	288.45	290.05
NKE	Nike Inc.	101.98	76.85	76.59
PG	Procter & Gamble Co.	169.06	171.35	168.76
KO	Coca-Cola Co.	62.17	63.84	63.35
PEP	PepsiCo Inc.	171.53	154.80	150.27
WMT	Walmart Inc.	88.76	96.82	99.54
COST	Costco Wholesale Corp.	711.12	924.35	1005.83
XOM	ExxonMobil Corp.	107.77	115.43	107.09
CVX	Chevron Corp.	150.98	153.45	149.34
COP	ConocoPhillips	98.46	103.50	98.48
SLB	Schlumberger Ltd.	40.05	38.54	40.16
MPC	Marathon Petroleum Corp.	139.70	160.12	147.00
BA	Boeing Co.	154.79	179.54	175.87
CAT	Caterpillar Inc.	362.51	385.42	361.55
UNP	Union Pacific Corp.	228.04	240.35	243.02
HON	Honeywell International Inc.	225.63	223.45	222.41
GE	General Electric Co.	165.26	169.45	204.11
DOW	Dow Inc.	40.09	52.85	37.38
NEM	Newmont Corp.	37.21	36.85	43.12
FCX	Freeport-McMoRan Inc.	38.03	39.45	35.86
LIN	Linde plc	417.67	415.32	454.21
ALB	Albemarle Corp.	86.44	85.32	80.45
DUK	Duke Energy Corp.	107.66	106.82	113.20
D	Dominion Energy Inc.	53.86	56.47	56.32
NEE	NextEra Energy Inc.	71.75	73.45	71.05
SO	Southern Co.	92.36	85.47	83.97
ED	Consolidated Edison Inc.	89.13	88.45	95.62
SPG	Simon Property Group Inc.	172.16	174.50	173.41
AMT	American Tower Corp.	183.25	178.50	185.37
PLD	Prologis Inc.	105.83	102.45	117.84
GOOG	Alphabet Inc.	190.58	188.76	202.64
META	Meta Platforms Inc.	585.46	69.67	697.46

Note: The price is indicated in USD.

APPENDIX C. MODEL PERFORMANCE VISUALIZATION

The bar chart below illustrates the comparative performance of GPT-4o and Claude 3.5 Sonnet based on five key forecast accuracy metrics: MAE, RMSE, MAPE, Pearson correlation coefficient (r), and R-squared (R²).

Figure C.1. Comparative performance of GPT-4o and Claude 3.5 Sonnet



Note: Claude 3.5 Sonnet demonstrated stronger performance across all evaluated metrics, with lower error rates and higher correlation to actual stock prices. This visualization supports the quantitative findings presented in Section 4.

APPENDIX D. SECTOR-LEVEL MODEL PERFORMANCE SUMMARY (MAPE COMPARISON)

The table below presents a comparison of forecasting accuracy, measured by MAPE, for GPT-4o and Claude 3.5 Sonnet across ten major industry sectors. The model with the lower MAPE in each sector is designated as the better-performing model.

Table D.1. A comparison of forecasting accuracy

Sector	*GPT-4o MAPE (%)*	*Claude 3.5 Sonnet MAPE (%)*	*Better model*
Consumer discretionary	11.29	3.17	Claude 3.5 Sonnet
Consumer staples	11.26	3.23	Claude 3.5 Sonnet
Energy	1.40	5.72	GPT-4o
Financials	29.60	24.18	Claude 3.5 Sonnet
Healthcare	7.99	5.99	Claude 3.5 Sonnet
Industrials	7.78	5.45	Claude 3.5 Sonnet
Materials	8.50	16.11	GPT-4o
Real estate	4.02	5.80	GPT-4o
Technology	8.09	12.09	GPT-4o
Utilities	5.41	3.71	Claude 3.5 Sonnet

Note: Model performance summary: Claude 3.5 Sonnet demonstrated superior predictive accuracy in 6 out of 10 sectors; GPT-4o outperformed Claude 3.5 Sonnet in 4 out of 10 sectors. Lower values indicate greater forecasting accuracy.

APPENDIX E. STATISTICAL METRICS AND FORMULAS USED IN MODEL EVALUATION

1. *MAE*: Measures the average magnitude of prediction errors without considering direction.

$$MAE = \left(\frac{1}{n}\right) * \sum |y_i - \hat{y}_i| \tag{1}$$

2. *RMSE:* Emphasizes larger errors due to squaring the residuals before averaging.

$$RMSE = \sqrt{\left[\left(\frac{1}{n}\right) * \sum (y_i - \hat{y}_i)^2\right]} \tag{2}$$

3. *MAPE*: Normalizes errors by actual values and expresses the result as a percentage.

$$MAPE = \left(\frac{100\%}{n}\right) * \sum \left|\frac{(y_i - \hat{y}_i)}{y_i}\right| \tag{3}$$

where y_i = 0 were excluded to avoid division by zero.

4. *Bias*: Assesses systematic over- or underestimation of stock prices.

$$Bias = \left(\frac{1}{n}\right) * \sum (\hat{y}_i - y_i) \tag{4}$$

5. *Pearson correlation coefficient (r)*: Measures the strength and direction of the linear relationship between actual and predicted values.

$$r = \sum \frac{[(y_i - (\bar{y})(\hat{y}_i - \hat{\bar{y}})]}{\left[\sqrt{\Sigma(y_i - \overline{y)}^2} * \sqrt{\Sigma(\hat{y}_i - \hat{\bar{y}})^2}\right]} \tag{5}$$

6. *Coefficient of determination (R²):* Indicates the proportion of variance in actual values explained by the model predictions.

$$R^2 = 1 - \left[\frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}\right] \tag{6}$$