SESSION: CORPORATE GOVERNANCE AND AI

LEVERAGING ARTIFICIAL INTELLIGENCE MODELS FOR FINANCIAL FORECASTING

Mfon Akpan *

* Northeastern State University, Tahlequah, USA

How to cite: Akpan, M. (2025). Leveraging artificial intelligence models for financial forecasting. In A. M. Gallo, U. Comite, & A. Kostyuk (Eds.), *Corporate governance: International outlook* (pp. 17–23). Virtus Interpress. https://doi.org/10.22495/cgiop3

Received: 31.03.2025
Accepted: 05.05.2025
Keywords: AI Forecasting, GPT-4o, Claude Sonnet 3.5, Financial Modeling, Stock Prediction, Machine Learning
JEL Classification: G17, C53, C45, G11
DOI: 10.22495/cgiop3

Abstract

This study investigates the predictive validity of generative artificial intelligence (AI) in financial forecasting. Specifically, it evaluates the zero-shot forecasting capabilities of GPT-4o and Claude Sonnet 3.5 by comparing their predicted stock prices against actual closing prices from a cross-industry portfolio as of February 3, 2025. Utilizing standardized statistical measures such as mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), correlation coefficients, and R^2, the study finds that Claude Sonnet 3.5 consistently outperforms GPT-4o in predictive accuracy and correlation. The research also examines directional bias and sector-specific performance.

1. INTRODUCTION

1.1. Purpose of study

This study evaluates the zero-shot forecasting capabilities of large language models (LLMs) in predicting stock prices. It compares GPT-4o

and Claude Sonnet 3.5 concerning the actual closing prices of a multi-sector portfolio as a measure of predictive precision, bias, and statistical reliability (Sangeetha & Alfia, 2024; El-Azab et al., 2024).

1.2. Importance of artificial intelligence (AI) in finance

AI integration in capital markets marks a shift in financial analytics, especially predictive modeling. Traditional methods, such as technical and fundamental analysis, struggle with asset prices' stochastic, non-stationary nature (Billah et al., 2024). Generative AI models offer high-frequency, autonomous forecasting that may enhance or surpass conventional techniques (Buczyński et al., 2023).

1.3. Why GPT-4o and Claude 3.5 are studied?

GPT-4o and Claude Sonnet 3.5 are the pioneering generative LLMs that harness deep learning and natural language processing skills to amalgamate intricate financial signals. They offer zero-shot prediction capabilities that need not be fine-tuned and thus become prime contenders in verifying empirical validity when markets turn volatile (Lin & Lobo Marques, 2024; Thrun, 2022).

1.4. Research gap and objectives

Existing literature shows the growing influence of AI models in financial forecasting and provides insight into the complexity and variability of their outcomes. While AI systems, including LLMs, offer enhanced analytical capacity, their performance is contingent upon input quality, model calibration, and contextual alignment (Chen et al., 2023; Bahoo et al., 2024). There is a recognized gap in empirical studies examining the zero-shot predictive performance of advanced LLMs such as GPT-4o and Claude Sonnet 3.5 across a multi-sectoral portfolio. This narrow scope limits understanding of how model behavior shifts with sector-specific volatility, firm size, or data characteristics.

This study fills that gap by evaluating two advanced LLMs, GPT-4o and Claude Sonnet 3.5, under zero-shot conditions, where no domain-specific training is applied. Predictions are made with standard prompts and evaluated across a diversified portfolio of more than 50 companies across ten large industries. It presents a comparative performance benchmark for LLMs in equity forecasting by measuring prediction accuracy, directional bias, and correlation with accurate prices.

2. METHODOLOGY

2.1. 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, OpenAI's GPT-4o and Anthropic's Claude Sonnet 3.5, 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.

2.2. Data collection procedures

Historical stock data for 2024 were retrieved from Google Finance using Google Sheets and Apps Script for transparency and reproducibility. Forecasts were generated from GPT-4o and Claude Sonnet 3.5 using standardized zero-shot prompts, ensuring consistency and eliminating user bias.

2.3. 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 from the following industries: technology, healthcare, financial services, consumer discretionary, consumer staples, energy, industrials, materials, utilities, and real estate.

2.4. 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:

- mean absolute error (MAE);
- root mean squared error (RMSE);
- mean absolute percentage error (MAPE);
- bias analysis;
- Pearson correlation coefficient (r);
- linear regression analysis (R^2).

2.5. 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.

2.6. 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 Sonnet 3.5, generated zero-shot stock price forecasts based on standardized prompts. Analytical computations were performed using a large language model (OpenAI's ChatGPT), executing Python-based analysis with pandas, NumPy, and scikit-learn. The following statistical measures were calculated for each model: MAE, RMSE, MAPE, bias, Pearson correlation coefficient (r), and coefficient of determination (R^2). 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 and ethical considerations were adhered to.

3. RESULTS

This section presents the empirical findings from the head-to-head comparison between GPT-4o and Claude Sonnet 3.5, 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 1. Model Performance Metrics

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

Note: All metrics include outlier data. Predictions for high-value, high-volatility firms, such as Goldman Sachs (GS) and Costco (COST), resulted in significant errors for both models.

The results may be interpreted as follows:

- Claude Sonnet 3.5 outperformed across most metrics, with lower MAE and RMSE, indicating forecasts closer to actual prices.
- Claude Sonnet 3.5 had a lower bias (-15.53) than GPT-4o (-23.82), showing a more neutral error pattern, while GPT-4o tended to underpredict, especially for high-priced stocks.
- Both models showed strong correlation and R^2 values, with Claude Sonnet 3.5 slightly higher ($r = 0.9096$, $R^2 = 0.8208$), better aligning with market trends.
- GPT-4o had a slightly better MAPE, suggesting stronger percentage-based accuracy on lower-priced stocks, making it suitable for proportional accuracy needs.

4. DISCUSSION

The results from this comparative analysis contribute meaningful insights into the predictive capabilities of LLMs in stock market forecasting under zero-shot conditions.

This comparative analysis provides meaningful insight into how current-generation LLMs perform in stock market forecasting under zero-shot prompting conditions. The evaluation revealed that both Claude Sonnet 3.5 and GPT-4o possess measurable forecasting capabilities, though they differ in precision, consistency, and bias.

4.1. Claude Sonnet 3.5

Claude Sonnet 3.5 consistently outperformed GPT-4o across most performance metrics, including lower MAE (29.98), lower RMSE (82.82), and reduced bias (-15.53). This indicates that its predictions were closer to the actual stock prices and more directionally balanced. Its R-squared value ($R^2 = 0.8208$) and high correlation ($r = 0.9096$) show that Claude Sonnet 3.5 captured the underlying market trend more effectively and explained more of the variance in actual stock prices.

4.2. GPT-4o

GPT-4o produced slightly better MAPE (9.67%) than Claude Sonnet 3.5 (8.55%), suggesting relatively more substantial proportional accuracy on lower-priced stocks. However, its bias score (-23.82) reveals a consistent tendency to underpredict, particularly for large-cap equities. This cautious forecasting behavior could be advantageous in risk-sensitive settings but may limit market performance characterized by upward movement or aggressive price action. Its high correlation ($r = 0.8932$) confirms that GPT-4o successfully captured directional trends, though with less magnitude accuracy than Claude.

4.3. Outliers and real-world implications

Rather than remove extreme cases, this study intentionally included outliers to evaluate each model's robustness. Notably, the most significant prediction discrepancies appeared in financials (e.g., GS) and consumer staples (e.g., COST). These outliers significantly inflated RMSE but mirror real-world forecasting conditions where unpredictability is inevitable. Claude's performance remained superior even under these conditions.

5. CONCLUSION

This study compared the zero-shot forecasting performance of GPT-4o and Claude Sonnet 3.5 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 Sonnet 3.5 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. Both models strongly correlate with actual prices.

This study has several limitations that should be acknowledged:

- The analysis represents a single forecasting window with a static prediction date (February 3, 2025).
- All forecasts were generated using a zero-shot prompt without domain-specific tuning or time-series modeling enhancements.
- The models did not ingest actual CSV files but inferred historical trends based on prompt structure, highlighting an area for deeper integration in future studies.

This study proposes the following recommendations:

1.Deploy Claude Sonnet 3.5 in financial decision systems.

2.Claude's stronger performance across most error metrics suggests it is well-suited for portfolio modeling, risk forecasting, and trading analytics applications.

3.Apply sector-specific prompt calibration.

4.Sector-level MAPE analysis revealed performance differences. Tailoring prompts to industry context (e.g., volatility or seasonality) may enhance precision.

5.Implement outlier detection and monitoring.

6.Real-world deployments should include error thresholding systems to flag extreme or unexpected predictions for manual review.

7.Establish model drift monitoring.

8.LLM performance should be audited regularly as market dynamics evolve. Prompt engineering and retraining cycles should be part of ongoing model governance.

9.Ensure compliance and explainability.

10.AI-generated forecasts must include transparency and traceability mechanisms, especially in regulated banking or asset management sectors.

11.Invest in AI-human collaboration models.

12.Generative AI is best used as an augmentation tool. Financial teams should consider frameworks where LLMs assist analysts with structured outputs rather than replacing expert judgment.

REFERENCES

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

Billah, M. M., Sultana, A., Bhuiyan, F., & Kaosar, M. G. (2024). Stock price prediction: Comparison of different moving average techniques using a 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

Chen, Z., Balan, M. M., & Brown, K. (2023). *Language models are few-shot learners for prognostic prediction*. ArXiv. https://arxiv.org/abs/2302.12692

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

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

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

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