



Article

Evaluating the Efficacy of Large Language Models in Stock Market Decision-Making: A Decision-Focused, Price-Only, Multi-Country Analysis Using Historical Price Data

Maria C. Mariani ^{1,2,†}, Sourav Malakar ^{2,*,†} , Amrita Bagchi ^{3,†}, Subhrajyoti Basu ^{4,†}, Saptarsi Goswami ^{5,†}, Osei Kofi Tweneboah ^{6,†} , Sarbadeep Biswas ^{7,†}, Ankit Dey ^{4,†} and Ankit Sinha ^{8,†}

¹ Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968, USA; mcmariani@utep.edu

² Department of CSE, Institute of Engineering & Management (IEM), Kolakta 700091, India

³ Society of Data Science, Pune 411061, India; amritabagchi3@gmail.com

⁴ Department of CSE, Heritage Institute of Technology, Kolkata 700107, India; subhrajyoti479@gmail.com (S.B.)

⁵ Department of Computer Science, Bangabasi Morning College, Kolkata 700009, India

⁶ Ramapo Data Science Program, Ramapo College of New Jersey, Mahwah, NJ 07430-1680, USA

⁷ Department of CSE, Adamas University, Kolkata 700126, India; sarbadeep.biswas@stu.adamasuniversity.ac.in

⁸ Department of CSE, Flame University, Pune 412115, India

* Correspondence: sourav.malakar@iem.edu.in

† These authors contributed equally to this work.

Abstract

This study provides a comparative evaluation of three state-of-the-art large language models (LLMs), namely OpenAI's (San Francisco, CA, USA) GPT-4.0, Google's (Google LLC, Mountain View, CA, USA) Gemini 2.0 Flash, and Meta's (Meta Platforms, Menlo Park, CA, USA) LLaMA-4-Scout-17B-16E, in a decision-oriented framework in which the models generate structured outputs based only on historical closing-price data. The evaluation covers 150 stocks sampled from three countries (India, the United States, and South Africa) across ten economic sectors, including Information Technology, Banking, and Pharmaceuticals. Unlike many prior studies that combine numerical and textual inputs, this study relies solely on three years of numerical time series data and examines model responses in terms of decision labels such as buy, sell, or hold. The LLMs were provided with historical closing-price sequences and prompted with three types of finance-related questions: (a) whether to buy a stock, (b) whether to sell or hold a stock, and (c) in a pairwise comparison, which stock to buy or hold. These prompts were evaluated across two investment horizons: 1 month and 3 months. Model outputs were compared against realized market outcomes during the corresponding test periods. Performance was assessed across four key dimensions: country, sector, annualized volatility, and question type. The models were not given any supplementary financial information or instructions on specific analytical methods. The results indicate that GPT-4.0 achieves the highest average accuracy (56%), followed by LLaMA-4-Scout-17B-16E (48%) and Gemini 2.0 Flash (39%). Overall performance remains moderate and varies across market conditions, with relatively higher accuracy observed in high-volatility regimes (51%). This work evaluates how LLMs behave when presented with structured numerical price sequences in a controlled decision-labeling setting and contributes to the broader discussion on the potential and limitations of LLMs for numerical decision tasks in finance.

Keywords: stock market; LLM; stock price decision making



Academic Editor: Oliver Hinz

Received: 8 February 2026

Revised: 1 April 2026

Accepted: 4 April 2026

Published: 17 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Recent advances in large language models (LLMs)—such as OpenAI’s GPT-4.0 [1], Google’s Gemini 2.0 Flash [2], and Meta’s LLaMA-4-Scout-17B-16E [3]—have significantly expanded the boundaries of natural language understanding and a wide range of prompt-based tasks. These models have demonstrated strong performance in a wide array of tasks, including question answering [4,5], sentiment analysis [6,7], summarization [8], and even vision–language understanding. In contrast, their application to structured, time series prediction—particularly in quantitative finance [9] and stock market [10–13] forecasting based on direct numerical data—remains relatively less developed. While initial works [14–16] have begun to explore LLMs for financial forecasting and general time series modeling [17,18], these approaches are still limited in scope compared to the rich body of NLP-focused applications.

Traditional approaches to financial time series modeling have relied on statistical methods such as ARIMA and GARCH [4,19], which are effective for linear patterns but often fail to capture nonlinear dependencies, level shifts, and structural breaks. With the rise of deep learning architectures, particularly LSTMs [20] and Transformers, more complex temporal dynamics could be modeled, yielding improved predictive accuracy. However, these architectures typically require task-specific model design, careful hyperparameter tuning, and significant training resources. Furthermore, their outputs often remain narrowly predictive (e.g., future price estimates) rather than decision-oriented, making them less directly useful to investors.

By contrast, LLMs offer a new paradigm; their generalist training allows them to process diverse forms of input, including structured numerical sequences presented through prompting—including, in principle, structured numerical sequences—without requiring extensive retraining. This raises an intriguing possibility—can LLMs be repurposed for decision-oriented financial tasks, where the output is not merely a predicted price but a discrete decision label such as “buy,” “hold,” or “sell”? Such a shift would align model evaluation more closely with decision-oriented evaluation settings, where interpretability or raw prediction accuracy may be secondary to the quality of discrete actions. It is important to note that the models are not used for analyzing investor reasoning or behavioral finance, but purely as quantitative forecasting architectures applied to closing-price time series data. The use of LLMs in this context refers to their computational structure and learning capability rather than linguistic or interpretive analysis.

Early explorations have begun to test this hypothesis. Ref. [14] demonstrated that LLMs can be reprogrammed to handle time series inputs through tokenization strategies, while ref. [15] proposed prompt-based methods for representing stock data as textual sequences suitable for LLM consumption. Similarly, ref. [21] studied LLMs in financial prediction but remained focused primarily on text-based sentiment and news-driven signals rather than direct numerical series. Collectively, these works suggest that LLMs are capable of engaging with financial time series problems, but evaluations remain narrow—often restricted to a handful of stocks, a single market, or a limited set of tasks. Importantly, few studies have systematically tested LLMs on direct decision-making from raw historical numerical data across diverse markets and industries, leaving open the question of how well their behavior under structured numerical prompting extends to comparable forecasting tasks.

This study aims to address this gap by designing an experimental framework in which LLMs are evaluated on structured stock decision tasks using only historical stock closing prices. The evaluation covers 150 stocks across three countries—India, South Africa, and the United States—and spans ten industry sectors representing a range of market capitalization and volatility profiles. To capture different decision contexts, we employ three types of

investor-relevant prompts (buy, sell/hold, and pairwise comparison) across two investment horizons (1 month and 3 months). This setup enables a structured assessment of how well LLMs can transform raw sequential numerical information into structured decision outputs under a common evaluation framework.

Although the theory of finance claims that short-term stock prices follow a random walk according to Efficient Market theory, this work is not designed to test the hypothesis of Efficient Market theory. This work is designed to test the behavior of modern LLMs when provided with only historical stock prices in terms of their consistency and relative accuracy under the study's decision-labeling framework. Prior work [22] shows that price movements can be described through statistical patterns. In a similar spirit, this study adopts a data-driven approach. This study is not intended as a head-to-head benchmark against trained forecasting systems; rather, it evaluates the relative behavior of general-purpose LLMs under a common zero-shot, price-only decision framework. To provide an initial economic perspective on the decision outputs, we also include a simple Sharpe ratio check, while noting that the study does not constitute a full trading strategy evaluation.

The main contributions of this study are listed below:

- It expands the scope of LLM assessment beyond language-centric tasks to structured numerical input, establishing a framework for decision-focused evaluation.
- This study examines LLM behavior under structured numerical input and frames evaluation around standardized decision labels.
- To the best of our knowledge, this is among the first large-scale, cross-country evaluations of LLMs on structured, price-only time series data, focused explicitly on generating decision labels such as buy, sell, or hold.
- It highlights the opportunities and limitations of using LLMs under a common prompt-based evaluation setting, offering insights into their comparative behavior in domains traditionally dominated by statistical and deep learning models.

The rest of this paper is organized as follows. Section 2 presents a detailed review of the current literature on LLMs and financial forecasting. It highlights important studies, gaps in methods, and the reasons for this research. Section 3 describes the data collection process, prompts, dataset preparation, details of the model implementation, and evaluation framework. Section 4 presents the empirical results and discussion. Section 5 concludes with key findings, limitations, and suggestions for future research.

2. Literature Review

2.1. Classical and Deep Learning-Based Feature Engineering in Financial Forecasting

Financial forecasting has long relied on statistical and machine learning methods [23], ranging from econometric models such as ARIMA [4,24,25] and GARCH [19,26,27] to neural architectures such as LSTMs [28,29], GRUs [30], and Transformers [20]. While these methods have achieved notable success in capturing temporal dependencies, they often require domain-specific feature engineering [31], careful hyperparameter tuning, and remain sensitive to data availability and market volatility. Beyond purely time series data, researchers have also incorporated broader firm-level factors, such as CEO power and labor productivity [32,33], further extending the scope of predictive models. Hybrid models, such as the LSTM-mTrans-MLP, integrate recurrent, Transformer, and feed-forward architectures to enhance robustness and accuracy across datasets like Bitcoin, CSI 300, and S&P 500 constituents [34]. Comparative evaluations consistently show that Transformer-based models outperform recurrent networks in capturing long-range dependencies in volatile markets.

2.2. LLMs in Financial Forecasting

Recent advances in LLMs have opened a new line of research—whether models trained on vast textual corpora can be adapted to structured financial prediction tasks. The central idea in this emerging literature is to reformat numerical data into natural language prompts, thereby enabling LLMs to leverage pre-trained reasoning abilities in non-textual domains. For example, ref. [15] explored GPT-4.0 for stock movement prediction by converting historical price sequences into textual descriptions, while ref. [14] assessed prompt-based approaches for structured financial data. These studies demonstrated that LLMs can capture broad market trends but struggle with fine-grained temporal predictions, making them less suitable for high-stakes decision-making.

2.3. Multi-Modal and Hybrid LLMs in Financial Forecasting

Other works have pursued hybrid and multi-modal designs that combine text and numeric signals. Ref. [35] compared GPT-4.0 and Gemini 2.0 Flash for long-horizon, cross-sector forecasts. Similarly, ref. [36] introduced LLMFactor, which integrates financial theory with LLMs for interpretable stock movement prediction. More recently, ref. [37] presented StockLLM, a retrieval-augmented model that unifies textual and numeric inputs for financial forecasting, showing that while LLMs can synthesize heterogeneous data effectively, their generalization remains inconsistent across market contexts. Ref. [16] proposes and tests the use of LLMs for explainable financial time series forecasting. They show that LLMs can combine numerical price data and textual news to perform better than traditional models and offer reasoning that is easy for humans to understand. Ref. [38] introduces the RiskLabs framework to evaluate ChatGPT's effectiveness in financial forecasting, highlighting both its predictive potential and its limitations in real-world markets.

2.4. Retrieval-Augmented Generation (RAG) and LLMs in Financial Forecasting

Retrieval-augmented generation (RAG) [39] has emerged as a promising paradigm for extending these models. TimeRAG [40] applied dynamic time warping to retrieve relevant sequences, while RAF [41] generalized the RAG concept to time series foundation models. Beyond numeric prediction, RAG-enhanced approaches have also been applied to financial sentiment analysis, where retrieval improves the robustness of LLM-generated predictions [42]. Complementary lines of research integrate LLMs with domain-specific heuristics: for example, ElliottAgents [43] combine Elliott Wave theory, RAG, and reinforcement learning in a multi-agent framework for technical analysis, demonstrating how LLMs can complement rather than replace expert-driven methods. Ref. [44] introduces a new retrieval-augmented time series diffusion model (RATD). This model combines an embedding-based retrieval method with a reference-guided denoising process. Together, these features greatly enhance forecasting accuracy for complex time series tasks.

2.5. Research Gaps and Questions

Despite these advances, several limitations persist. First, most prior studies have focused narrowly on the US market, with limited attention to cross-country or cross-sector validation. Second, many approaches rely on hybrid text-numeric inputs (e.g., reports + prices), leaving the question of whether LLMs can operate effectively on purely numerical time series data underexplored. Third, while RAG frameworks improve plausibility, their quantitative rigor and consistency remain below specialized econometric and deep learning models. Finally, the challenges of reproducibility, interpretability, and robustness under diverse market conditions remain unresolved.

In summary, the existing literature provides encouraging evidence that LLMs and RAG-based models can contribute to financial forecasting. However, systematic evaluations

of their ability to generate consistent, decision-oriented predictions (e.g., buy/sell/hold) across geographies, industries, and market capitalizations are still lacking. This study aims to address these gaps by evaluating leading LLMs on diverse, real-world stock datasets using only historical closing price data, thereby isolating their core decision behavior under structured numerical input.

To guide the analysis, we pose the following research questions (RQs):

- RQ1: To what extent can LLMs (GPT-4.0, Gemini 2.0 Flash, LLaMA-4-Scout-17B-16E) generate accurate decision labels for short-term (1-month) and medium-term (3-month) investment outcomes when provided only with historical closing price data?
- RQ2: How does decision performance vary with contextual factors such as stock volatility, industry sector, and country of listing?
- RQ3: Which LLM demonstrates the greatest consistency and relative accuracy across different evaluation settings?
- RQ4: How does the type of investment question—buy, sell/hold, or pairwise comparison—affect model accuracy and reliability?

3. Materials and Methods

This section describes the critical aspects of the experiment.

3.1. Stock Identification and Closing-Price Collection

Stocks were identified from three countries, namely United States of America, South Africa, and India. The United States, with over 5000 listed equities and a total market capitalization of approximately USD 40 trillion, represents a mature and highly liquid market characterized by significant institutional participation. India is an emerging market with high growth potential and has a significant number of active retail investors, while South Africa is considered a frontier market with differing economic conditions and has significant commodity price links. Together, these three markets capture a diverse range of market efficiency levels, volatility patterns, and investor behaviors. Ten major industry sectors were considered: Financial, Energy, Consumer Staples, Healthcare, Technology, Real Estate, Materials, Consumer Discretionary, Industrials, and Communication Services [45]. A total of fifty stocks were selected, with five stocks drawn from each of the ten sectors. Only those stocks were included for which data were available for three years and which did not undergo dilution through stock splits, bonus issues, or similar corporate actions. This selection process yielded 50 stocks per country, resulting in a total of 150 stocks across all markets. The complete list of selected stocks is provided in Table 1.

Data sources were tailored to each market. For the US market, stock information was retrieved via Yahoo Finance (Yahoo Inc., Sunnyvale, CA, USA) [46]. For South Africa, data were downloaded from the Johannesburg Stock Exchange (JSE) [8]. For India, stock information was obtained from the National Stock Exchange of India (National Stock Exchange of India, Mumbai, India) [47]. For each stock, daily closing prices were collected for the period 1 March 2022, to 28 March 2025, and stored in structured comma separated value files. To ensure consistency and scalability, a suite of Python-based automation scripts was developed to systematically extract, clean, and consolidate historical stock data across markets and capitalization levels. This repository forms the foundation of the prediction dataset used in our experiments.

Table 1. Stocks by Sector and Country.

Sector	India	USA	South Africa
Communication Services	Bharti Airtel Limited, Bharti Hexacom Limited, Indus Towers Limited, Tata Communications Limited, Tejas Networks Limited	AT&T Inc., Comcast Corporation, Meta Platforms Inc., T-Mobile US Inc., Verizon Communications Inc.	Hudaco Industries Limited, MTN Group Limited, Telkom SA SOC Limited, Vodacom Group Limited
Consumer Discretionary	Blue Star Limited, Crompton Greaves Consumer Electricals Limited, Havells India Limited, Titan Company Limited, Whirlpool of India Limited	Amazon.com Inc., Ford Motor Company, Nike Inc., Tesla Inc., The Home Depot Inc.	African and Overseas Enterprises Limited, City Lodge Hotels Limited, Famous Brands Limited, Lewis Group Limited, Tsogo Sun Gaming Limited
Consumer Staples	Dabur India Limited, Emami Limited, Godrej Consumer Products Limited, Hindustan Unilever Limited, ITC Limited, Marico Limited, Nestle India Limited	Coca-Cola Company, Costco Wholesale Corporation, PepsiCo Inc., Procter & Gamble Company, Walmart Inc.	AH-Vest Limited, Crookes Brothers Limited, Pepkor Holdings Limited, RCL Foods Limited, Tiger Brands Limited
Energy	GAIL (India) Limited, Oil and Natural Gas Corporation Limited, Reliance Industries Limited, Tata Power Company Limited	Chevron Corporation, ConocoPhillips, Exxon Mobil Corporation, NextEra Energy Inc., Schlumberger Limited	Efora Energy Limited, Exxaro Resources Limited, Sasol Limited, Thungela Resources Limited, TotalEnergies Marketing South Africa (Pty) Ltd.
Financials	AU Small Finance Bank Limited, Bandhan Bank Limited, HDFC Bank Limited, ICICI Bank Limited, IndusInd Bank Limited, Kotak Mahindra Bank Limited, State Bank of India	Bank of America Corporation, Citigroup Inc., JPMorgan Chase & Co., Morgan Stanley, Wells Fargo & Company	African Dawn Capital Limited, Finbond Group Limited, Investec Limited, Nedbank Group Limited, Standard Bank Group Limited
Healthcare	Aurobindo Pharma Limited, Cipla Limited, Dr. Reddy's Laboratories Limited, Lupin Limited, Sun Pharmaceutical Industries Limited	Abbott Laboratories, Johnson & Johnson, Merck & Co. Inc., Pfizer Inc., UnitedHealth Group Incorporated	Advanced Call Center Technologies, Ascendis Health Limited, Aspen Pharmacare Holdings Limited, Dis-Chem Pharmacies Limited, Netcare Limited
Industrials	GMR Infrastructure & GVK Power and Infrastructure Limited, Larsen & Toubro Limited, Siemens Limited, Thermax Limited, Voltas Limited	3M Company, Boeing Company, Caterpillar Inc., General Electric Company, Honeywell International Inc.	Barloworld Limited, Brikor Limited, Calgro M3 Holdings Limited, Murray & Roberts Holdings Limited, Reunert Limited
Information Technology	Infosys Limited, L&T Technology Services Limited, Tata Consultancy Services Limited, Tech Mahindra Limited, Wipro Limited	Alphabet Inc., Apple Inc., International Business Machines Corporation, Microsoft Corporation, NVIDIA Corporation	AYO Technology Solutions Limited, Altron Limited, Datatec Limited, EOH Holdings Limited, Mustek Limited
Materials	Hindalco Industries Limited, Hindustan Zinc Limited, JSW Steel Limited, Steel Authority of India Limited, Tata Steel Limited	Alcoa Corporation, Freeport-McMoRan Inc., Nucor Corporation, The Mosaic Company, United States Steel Corporation	Afrimat Limited, Anglo American plc, ArcelorMittal South Africa Limited, Harmony Gold Mining Company Limited, Impala Platinum Holdings Limited
Real Estate	Brigade Enterprises Limited, DLF Limited, Godrej Properties Limited, Oberoi Realty Limited, Prestige Estates Projects Limited	AvalonBay Communities Inc., Equity Residential, Prologis Inc., Simon Property Group Inc., Welltower Inc.	Growthpoint Properties Limited, Putprop Limited, Redefine Properties Limited, Resilient REIT Limited, Vukile Property Fund Limited

3.2. Prompts

The investment tasks in this study are framed around three types of investor-relevant queries that capture common decision-making scenarios in financial markets:

- (a) Buy Decision—should the investor purchase a particular stock, given its price history? (Repeated for both 1 and 3 months.)
- (b) Sell/Hold Decision—should the investor sell a currently held stock, or continue to hold it? (Same investment horizon as buy.)
- (c) Comparison Decision—between two candidate stocks, which is a better investment choice over a specified horizon? (Same investment horizon as buy.)

To operationalize these tasks, we developed a set of sample prompts that standardize how queries are presented to the LLMs. Prompt patterns or templates are well suited to get results efficiently [48]. A constant system instruction was used for all experiments. For APIs that support system prompts, this was presented to the models as a system prompt. For other models, this instruction was presented within the beginning of the user prompt. The actual text of this instruction was identical for all models.

- System Prompt:
You are a financial analyst.
- Sample Buy Prompt:
Based on the daily closing price of stock [STOCK NAME] over the past [N] months (provided in a CSV file), should this stock be bought today for a [M]-month investment horizon? Please respond with Buy or Not to Buy.
- Sample Sell/Hold Prompt:
Given the daily closing prices of stock [STOCK NAME] over the last [N] months (provided in a CSV file), and assuming the stock is currently held in the portfolio, should it be sold at the end of the next [M] months or held beyond that period? Please respond with Sell or Hold.
- Sample Comparison Prompt (Buy):
Given the daily closing prices of two stocks—[STOCK A] and [STOCK B]—over the last [N] months (provided in two CSV files), which stock is more suitable to be bought today for a [M]-month investment horizon? Respond as None of them, Stock A, or Stock B.
- Sample Comparison Prompt (Sell/Hold):
Given the daily closing prices of two stocks—[STOCK A] and [STOCK B]—over the last [N] months (provided in two CSV files), which stock is more suitable to be sell or hold today for a [M]-month investment horizon? Respond as None of them, Stock A, or Stock B.

It may be noted, the data has been collected for March 2022–March 2025 (37 months). For the 1-month investment horizon, 36 months of data were used as input for the closing price. For the 3-month investment horizon, 34 months of data were used as input for the closing price, ensuring the prediction period does not overlap with the input period. More details on the prompts with sample data are provided in the Appendix A. The only available data for the models were the daily closing-price values. Other features like open, high, low, volume, and technical indicators were not provided. Each CSV input contained only two columns: Date and Close. The data was represented as plain text strings in the format of CSV. There were no pre-processing operations such as aggregation, truncation, and inclusion of calculated technical indicators. This ensures that the LLMs were used as purely numerical models. The length of each sequence used for input into the LLMs was approximately 730–780, entirely within the limits of the respective context window. There was no truncation for the API submission.

3.3. Methodology

In this study, the experimental setup adopts a multistep procedure to systematically evaluate the comparative performance of LLMs on structured decision-labeling tasks using

historical price data. The methodology, illustrated in Figure 1, ensures a reproducible workflow that includes data collection, prompt design, model query, and performance evaluation. Each component of the setup is described in detail below.

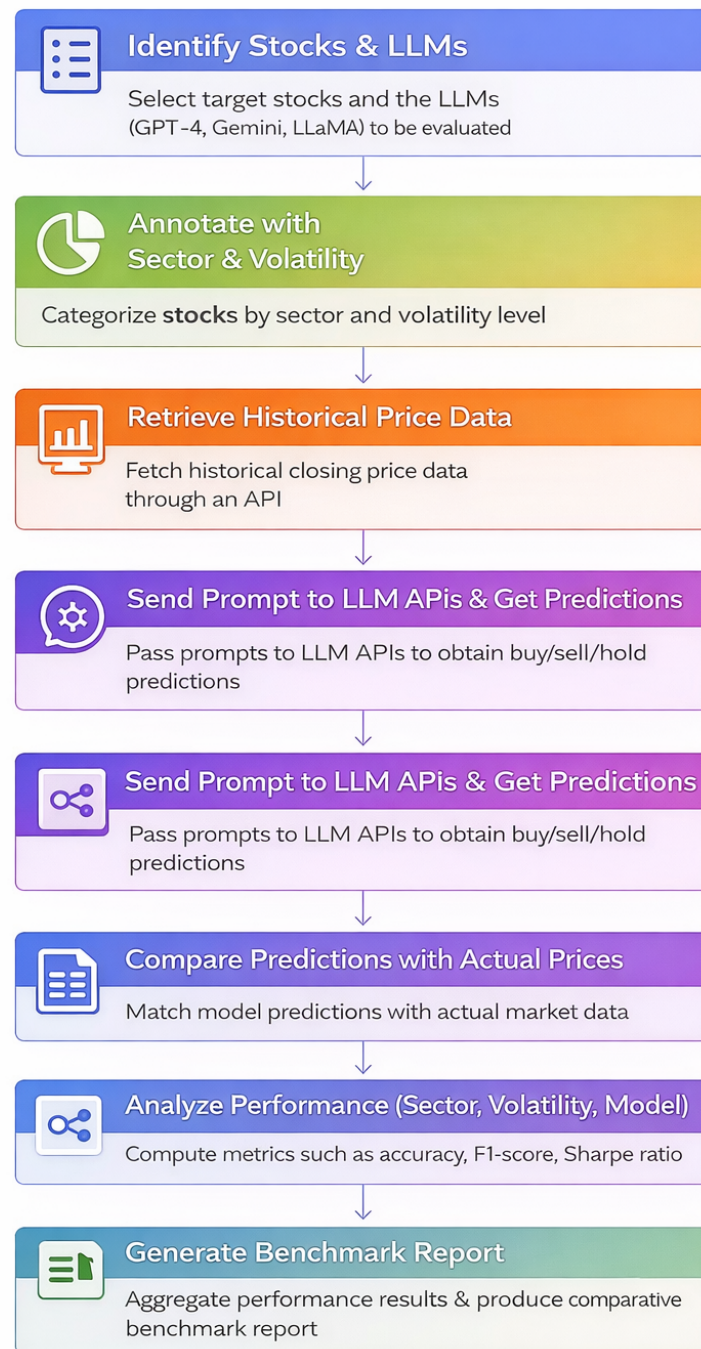


Figure 1. Methodology.

Step 1: Identify Stocks and LLMs: A representative set of 150 stocks was selected from three countries—India, the United States, and South Africa. These stocks span ten major economic sectors, ensuring coverage of diverse market sizes, sectoral dynamics, and volatility levels. Three state-of-the-art LLMs were chosen for evaluation—OpenAI’s GPT-4.0, Google’s Gemini 2.0 Flash, and Meta’s LLaMA-4-Scout-17B-16E—providing a range of architectures and training paradigms.

Step 2: Annotate with Sector and Volatility:

Each stock was annotated with its corresponding industry sector following the Global Industry Classification Standard (GICS). To further capture the individual risk characteristics, we calculated each stock's historical annualized volatility based on the dispersion of daily returns over the year preceding the prediction window. Daily returns were computed from closing prices using natural log returns for better statistical properties (especially for compounding) as follows:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \ln\left(\frac{P_t}{P_{t-1}}\right), \quad (1)$$

where P_t is the closing price at day t and P_{t-1} is the closing price at day $t - 1$.

The daily variability (sample standard deviation) of the daily returns was calculated as

$$\sigma_{\text{daily}} = \sqrt{\frac{1}{N-1} \sum_{t=1}^N (r_t - \bar{r})^2}, \quad (2)$$

where N is the number of trading days in the one-year window and $\bar{r}$ is the sample mean of the daily returns over that window.

The annualized volatility was then computed from the variability of daily returns using the conventional square-root-of-time rule, assuming 252 trading days per year.

$$\sigma_{\text{annual}} = \sigma_{\text{daily}} \times \sqrt{252}. \quad (3)$$

Stocks are then classified into three categories of volatility: low (0–5%), medium (5–10%), and high volatility stocks (>10%). These thresholds are used as practical within-study grouping cutoffs and should not be interpreted as universal market-wide definitions of low, medium, and high volatility. This method is also in line with the earlier literature, where relative volatility categorization is employed to study the differential behavior of returns for different volatility groups of stocks [49–51].

$$\text{Volatility Group} = \begin{cases} \text{Low,} & \text{if } \sigma_{\text{annual}} \leq 5, \\ \text{Medium,} & \text{if } 5 < \sigma_{\text{annual}} \leq 10, \\ \text{High,} & \text{if } \sigma_{\text{annual}} > 10. \end{cases}$$

This structured classification enabled comparative evaluation across both *sectoral dynamics* and *volatility levels*, thereby facilitating a deeper understanding of whether the model's predictive accuracy and decision behavior varied across stable, moderately volatile, and highly volatile market environments.

- Step 3:** Retrieve Historical Price Data via API: Daily closing prices were collected for all stocks from 1 March 2022, to 28 March 2025. Data sources included Yahoo Finance (US stocks), NSE India (Indian stocks), and the Johannesburg Stock Exchange (JSE) platform (South African stocks). Automated Python scripts were implemented to query these APIs and save each stock's time series in a dedicated CSV file.
- Step 4:** Prepare Prompts with CSV Data: For each stock, historical closing prices were formatted into structured CSV files that served as inputs for the LLMs. The prompts were designed to reflect three primary categories of investment decision-making tasks: (i) buy, (ii) sell/hold, and (iii) pairwise comparison. Each prompt specified the investment horizon (1 month or 3 months). To isolate the effect of numerical data, no market commentary or sentiment was included.
- Step 5:** Query LLM APIs and Collect Predictions: Prompts were sent to the specific LLM APIs using Python scripts. Identical prompts were given to all models under

controlled conditions. The temperature setting was fixed at zero to reduce random variation. The model outputs were not generated as free-form text; instead, they were limited to a specific set of responses, based on clear format rules included in the prompt design. For *buy-side* evaluations, each model was instructed to respond strictly with either “Buy” or “Not to Buy.” For *sell-side* tasks, the allowed responses were “Sell” or “Hold.” In *pairwise comparison* tasks, answers were restricted to “Stock A,” “Stock B,” or “None.”

Step 6: Compare Predictions with Actual Outcomes: Model predictions were evaluated against realized stock price movements using explicitly defined forward-return rules. This definition is used consistently for evaluating buy/sell outcomes. For each stock and prediction date t , the forward return over an investment horizon T (either 1 month or 3 months) was computed from close-to-close prices as

$$R_{t,T} = \frac{p_{t+T}^{\text{close}} - p_t^{\text{close}}}{p_t^{\text{close}}} \times 100\%. \quad (4)$$

In order to check the robustness of the results, we carried out a sensitivity check for the thresholds of $\pm 1\%$ and $\pm 5\%$, shown in Table 2. As expected, the accuracy of the models changes depending on the thresholds applied due to the change in the strictness of the labels and the class distribution. The ranking of the models remains the same for all the thresholds applied: GPT-4.0 > LLaMA-4-Scout-17B-16E > Gemini 2.0 Flash. These results suggest that the comparative ordering of the models is stable across the tested thresholds, although absolute performance depends on the threshold choice. The intermediate threshold of $\pm 2.5\%$ is used as the default, as it offers better and stable performance across models, without the noise sensitivity of $\pm 1\%$ or the reduced signal frequency of $\pm 5\%$.

Table 2. Threshold Sensitivity Analysis.

Threshold	GPT-4.0	LLaMA-4-Scout-17B-16E	Gemini 2.0 Flash
$\pm 1\%$	53.42%	45.11%	46.38%
$\pm 2.5\%$	56.35%	48.05%	38.94%
$\pm 5\%$	49.28%	40.67%	31.85%

The threshold $\theta = 2.5\%$ is used as a practical decision filter to exclude economically insignificant price movements and avoid labeling noise-driven fluctuations as actionable signals. In short-horizon settings, small price changes are often dominated by market microstructure effects and may not reflect meaningful investment opportunities. Therefore, the threshold is interpreted as a minimum return hurdle rather than a structural parameter.

Ground-truth labels were derived using decision thresholds of $\pm\theta$, as summarized below.

(a) Single-stock tasks

- Buy task:

$$\text{Label} = \begin{cases} \text{Buy,} & \text{if } R_{t,T} \geq \theta, \\ \text{Not to Buy,} & \text{otherwise.} \end{cases}$$

- Sell task:

$$\text{Label} = \begin{cases} \text{Sell,} & \text{if } R_{t,T} \leq -\theta, \\ \text{Hold,} & \text{otherwise.} \end{cases}$$

(b) Pairwise (comparison) tasks

Let $R_{1,t,T}$ and $R_{2,t,T}$ denote the forward returns of Stock 1 and Stock 2, respectively.

- Comparison Buy:

$$\text{Label} = \begin{cases} \text{Stock 1,} & \text{if } R_{1,t,T} > R_{2,t,T} \text{ and } R_{1,t,T} \geq \theta, \\ \text{Stock 2,} & \text{if } R_{2,t,T} > R_{1,t,T} \text{ and } R_{2,t,T} \geq \theta, \\ \text{None,} & \text{if } R_{1,t,T} < \theta \text{ and } R_{2,t,T} < \theta. \end{cases}$$

- Comparison Sell:

$$\text{Label} = \begin{cases} \text{Stock 1,} & \text{if } R_{1,t,T} \leq -\theta \text{ and } R_{2,t,T} > -\theta, \\ \text{Stock 2,} & \text{if } R_{2,t,T} \leq -\theta \text{ and } R_{1,t,T} > -\theta, \\ \text{None,} & \text{if } R_{1,t,T} > -\theta \text{ and } R_{2,t,T} > -\theta, \\ \text{Both,} & \text{if } R_{1,t,T} \leq -\theta \text{ and } R_{2,t,T} \leq -\theta. \end{cases}$$

Each model prediction was scored as a binary outcome (1 = correct, 0 = incorrect) by comparing the predicted action with the corresponding ground-truth label derived from these rules. This explicit evaluation framework supports transparency and reproducibility in evaluating model decisions across all task types. Because fixed thresholds generate different event frequencies across contexts, the resulting labels should be interpreted as evaluation constructs within this protocol rather than as universally comparable trading events.

Step 7: Analyze Accuracy by Sector, Volatility, and Model: Accuracy scores were aggregated and analyzed across dimensions such as sector, volatility group, investment horizon, and model type. This enabled within-model and cross-model comparisons. Additional analyses assessed stability and consistency of performance across financial environments.

Step 8: Compile Benchmark Report: The results were consolidated into a benchmark report, which included overall accuracy scores, sector- and volatility-based breakdowns, and comparative performance across investment question types. This benchmark provides a structured reference point for comparing LLM behavior across markets, sectors, and prompt types.

3.4. Dataset Preparation

Algorithm 1 was utilized to measure risk by determining the daily and annual stock closing-price volatility. In our study, the volatility classification thresholds are set as $\tau_1 = 5$ and $\tau_2 = 10$, and, based on that, the stocks are classified as low-, medium-, and high-volatility stocks.

Table 3 describes the overview of the attributes of the dataset. It includes stock identifiers, sector classifications, volatility measures, model predictions, and evaluation flags. This structure promotes transparency and makes it easier to repeat the evaluation of LLM performance in different countries, sectors, and types of investment decisions.

Algorithm 1 Volatility Estimation from Daily Closing Prices

```

1: Input:
   Daily closing prices  $P_1, P_2, \dots, P_n$ 
   Trading days per year  $D$  (default  $D \leftarrow 252$ )
   Optional rolling window size  $w$  (days)
   Volatility thresholds  $\tau_1, \tau_2$  (default  $\tau_1 = 5, \tau_2 = 10$ )
2: Output: Daily volatility  $\sigma_{\text{daily}}$ , annual volatility  $\sigma_{\text{annual}}$ , volatility percentage  $\sigma_{\%}$ , volatility class
3: Sort prices in chronological order
4: Remove missing or invalid price entries
5:  $n \leftarrow$  number of remaining prices
6: if  $n < 2$  then
7:   return error (insufficient data)
8: end if
9: if rolling window  $w$  is provided then
10:  for each window segment of length  $w$  (sliding by 1 day) do
11:    Apply steps 11–18 to the prices inside the window
12:    Record the windowed outputs
13:  end for
14:  return windowed  $\sigma_{\text{daily}}, \sigma_{\text{annual}}, \sigma_{\%}, \text{class}$ 
15: end if
16: for  $t \leftarrow 2$  to  $n$  do
17:  Compute log-return:

```

$$r_t \leftarrow \ln\left(\frac{P_t}{P_{t-1}}\right)$$

```

18: end for
19: Number of returns:  $m \leftarrow n - 1$ 
20: Compute mean return:

```

$$\bar{r} \leftarrow \frac{1}{m} \sum_{t=2}^n r_t$$

```

21: Compute sample variance:

```

$$v \leftarrow \frac{1}{m-1} \sum_{t=2}^n (r_t - \bar{r})^2$$

```

22: Daily volatility:

```

$$\sigma_{\text{daily}} \leftarrow \sqrt{v}$$

```

23: Annual volatility:

```

$$\sigma_{\text{annual}} \leftarrow \sigma_{\text{daily}} \cdot \sqrt{D}$$

```

24: Volatility percentage:

```

$$\sigma_{\%} \leftarrow 100 \times \sigma_{\text{annual}}$$

```

25: Classify volatility:

```

$$\text{class} \leftarrow \begin{cases} \text{Low} & \text{if } \sigma_{\%} \leq \tau_1 \\ \text{Medium} & \text{if } \tau_1 < \sigma_{\%} \leq \tau_2 \\ \text{High} & \text{if } \sigma_{\%} > \tau_2 \end{cases}$$

```

26: return  $\sigma_{\text{daily}}, \sigma_{\text{annual}}, \sigma_{\%}, \text{class}$ 

```

Table 3. Detailed Description of Dataset Attributes.

Attribute	Detailed Description
Class Level	Defines the investment horizon (1-month or 3-month) and the type of action (buy, sell, hold, or pairwise comparison).
Country	The country where the stock is listed and traded (India, United States, or South Africa).
Sector 1	Primary industry sector of Stock 1, categorized using GICS (e.g., Healthcare, Financials, IT). Allows sector-wise performance analysis.
Ticker 1	The official trading symbol of Stock 1 (e.g., HDFCBANK.BSE, ICICIBANK.BSE, INDUSINDBK.BSE, AUBANK.BSE, etc.). Acts as a unique identifier.
Stock 1	The full company name of Stock 1 (e.g., HDFC Bank Limited, ICICI Bank Limited, IndusInd Bank Limited, AU Small Finance Bank Limited, etc.), mapped to its ticker.
Sector 2	Industry sector of Stock 2 (if applicable in pairwise comparisons). Enables cross-sector decision analysis.
Ticker 2	Trading symbol of Stock 2 (in pairwise comparison tasks).
Stock 2	Full company name of Stock 2, used in comparative scenarios.
Annual Volatility Stock 1	Yearly volatility (%) of Stock 1, calculated from daily returns, based on Algorithm 1. Quantifies risk levels for Stock 1.
Annual Volatility Stock 2	Yearly volatility (%) of Stock 2 (in comparison tasks). Assesses risk-return tradeoffs across alternatives.
Input	Refers to data source (Yahoo Finance, NSE India, Johannesburg Stock Exchange (JSE)) and period of analysis (March 2022–March 2025). Ensures reproducibility.
LLaMA-4-Scout-17B-16E	Categorical recommendation of Meta’s LLaMA-4-Scout-17B-16E model (Buy, Sell, Hold, Stock A, Stock B, None).
Gemini 2.0 Flash	Recommendation produced by Google’s Gemini 2.0 Flash model under identical conditions.
GPT 4.0	Recommendation produced by OpenAI’s GPT 4.0 model. Typically the strongest performer, though with higher variability.
Actual	Ground-truth investment action derived from realized forward price movement (e.g., Buy if returns are positive). Serves as benchmark.
LLaMA-4-Scout-17B-16E Flag	Binary indicator (1/0) showing whether LLaMA-4-Scout-17B-16E’s recommendation matched the actual outcome. Used for accuracy computation.
Gemini 2.0 FlashFlag	Binary indicator of Gemini 2.0 Flash’s correctness (1 = match, 0 = mismatch).
GPT-4.0 Flag	Binary indicator of GPT 4.0’s correctness compared to the actual decision.
Sum	Total number of models (out of 3) that correctly predicted the actual decision. Captures model consensus strength.

3.5. Experimental Setup

In this work, we verified three well-known LLMs: GPT-4.0 [1], Gemini 2.0 Flash-2.0 Flash [2], and LLaMA-4-Scout-17B-16E -4-Scout-17B-16E [52]. The choice of these three models was based on their diversity, representation from different providers, and the availability of stable API implementations during the experimental period. GPT-4.0, Gemini 2.0 Flash, and LLaMA-4-Scout-17B-16E are leading foundation models from OpenAI, Google, and Meta, covering both proprietary and open-weight approaches. The goal of this study is not to thoroughly assess all available LLMs but to evaluate key state-of-the-art architectures within a uniform zero-shot, price-only framework. Future research may broaden this evaluation to include more emerging models like DeepSeek and other foundation architectures. Their architectural details are summarized in Table 4. All experiments were

carried out in Python (version 3.12), using NumPy (version 2.1) [53] and Pandas (version 2.2) [54] for data handling, Matplotlib (version 3.9) and Seaborn (version 0.13) [55] for visualization, and SciPy [56] for statistical testing. We executed model queries through the OpenAI API, Google Generative AI SDK, and Hugging Face Transformers. This approach ensured reproducibility and consistent evaluation across different countries, sectors, and volatility levels.

Table 4. Comparison of LLM Architectures.

LLM	Parameters	Context Window	Multimodality Support
GPT-4.0	Unknown	128,000 tokens	Yes
Gemini 2.0 Flash-2.0 Flash	Unknown	2,000,000 tokens	Yes
LLaMA-4-Scout-17B-16E	17 B	120,000 tokens	No

3.6. Evaluation Metric

For consistency across tasks and markets, model performance is evaluated using **classification accuracy**, defined as the proportion of correct predictions over the total number of predictions as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}} \quad (5)$$

In our setting, a prediction is considered correct if the model's suggested decision (buy, sell, hold, or pairwise choice) matches the ground-truth label derived from the realized stock movement over the specified investment horizon (1 month or 3 months).

In addition, to provide a more comprehensive evaluation under potential class imbalance, we also calculated balanced accuracy, along with class-specific precision and recall for each decision category.

In addition to classification-based evaluation, an additional performance perspective on the model outputs is provided using a simple Sharpe ratio-based check. The trading signal generated using the predictions is combined with the returns generated in the future to calculate the strategy returns. The performance is measured using the Sharpe ratio as follows:

$$S = \frac{\mathbb{E}[R_t - R_f]}{\sigma(R_t)} \quad (6)$$

where R_t denotes the strategy return at time t , R_f is the risk-free rate, and $\sigma(R_t)$ is the standard deviation of returns. This analysis provides an additional perspective on the model outputs, but it should not be interpreted as a full real-world implementation assessment.

4. Results

In the Results Section, we systematically attempt to answer the research questions listed. The datasets, experimental notebooks, and prompt templates used in this study are openly available for reproducibility (<https://github.com/Rangan2005/Evaluating-the-Efficacy-of-Large-Language-Models-in-Stock-Market-Decision-Making> (accessed on 10 January 2026)). Section 4 is organized as follows: We first examine overall decision-making performance across different investment horizons. Next, we examine how context varies by country, sector, and volatility levels. Finally, we compare model performance using repeated-measures statistical testing.

4.1. RQ1: How Accurately Can LLMs Generate Investment Decisions for Short-Term (1-Month) and Medium-Term (3-Month) Horizons Using Only Historical Closing Price Data?

In Table 5, the 1-month and 3-month horizons show very similar central tendencies, with nearly identical mean accuracies of 44.18% and 43.87%. The dispersion is a little higher for the 3-month horizon, with an SD of 17.96%, compared to the 1-month horizon, which has an SD of 16.60%. The Welch's t -test shows no significant difference between the horizons, with $t = 0.075$ and $p = 0.941$. While the Shapiro–Wilk tests indicate slight deviations from normality, with $p < 0.05$, the overall results suggest similar predictive performance across the investment horizons. Although GPT-4.0 shows slightly better short-term performance, the lack of a statistically significant difference indicates that predictive behavior stays generally consistent across different time frames.

Table 5. Descriptive Statistics for Short-Term (1-Month) vs. Medium-Term (3-Month) Predictions with t -test and Shapiro–Wilk.

Horizon	Mean	Median	SD	t (Welch)	p (t)	Shapiro-W	Shapiro- p
Short-Term (1-Month)	44.18%	39.02%	16.60%	0.075	0.941	0.927	0.021
Medium-Term (3-Month)	43.87%	40.57%	17.96%			0.936	0.038

Note: 150 stocks $\times$ 3 LLMs = 450 observations per horizon (1-month and 3-month).

As can be seen in Figure 2, the overall performance for the two horizons is similar, with only minor differences in the central tendency. However, it can also be observed that the 3-month horizon has a larger spread and larger extremes, implying more variability in the models' performance. In contrast, the 1-month horizon has a smaller spread, implying stable and consistent performance. This implies that there is more uncertainty when the horizons are longer, while there is more reliability when the horizons are shorter.

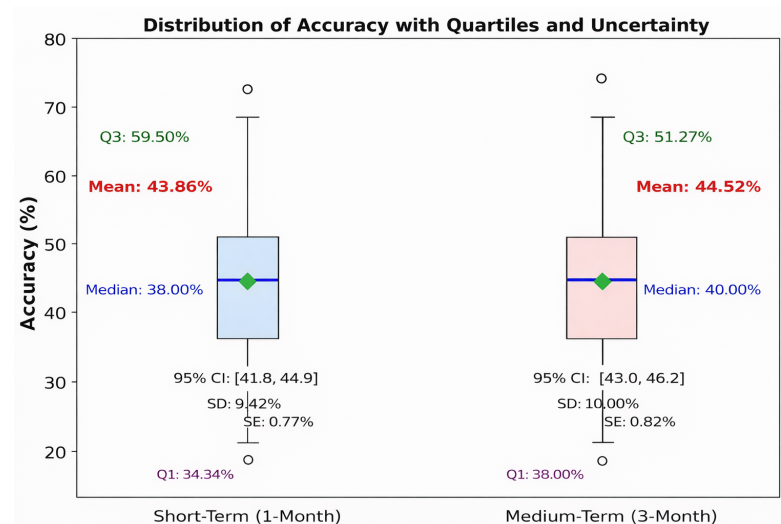


Figure 2. Distribution of model accuracy across 1-month and 3-month horizons. Boxplots display the interquartile range (Q1–Q3), medians (blue), and mean values (green diamonds). White circles represent outliers. Q1 and Q3 values are 34.34–50.50% for 1-month and 32.00–51.27% for 3-month horizons. Uncertainty is represented by 95% confidence intervals [42.63, 45.73] and [42.26, 45.48]), standard deviations (9.72%, 10.02%), and standard errors (0.79%, 0.82%).

In this sample, the shorter horizon shows slightly less variability than the longer horizon, as noted by [28]. The results suggest that short-term predictions may be more stable because they have less compounding uncertainty.

Although it is possible to get a baseline for the performance of a model through its classification accuracy, it does not capture the economic value of decisions generated by

a model. To provide an additional perspective on the model outputs, we also calculate a simple Sharpe ratio-based measure (described in Section 4.3).

4.2. RQ2: How Does Decision Performance Vary with Contextual Factors Such as Stock Volatility, Industry Sector, and Country of Listing?

4.2.1. Volatility-Level Analysis

As shown in Table 6, high volatility has the strongest central tendency, with a mean of 51.11% and a median of 51.31%. It also has tight dispersion, with a standard deviation of 8.69% and an inter-quartile range (IQR) of 45.18% to 58.68%. This suggests higher observed accuracy in the high-volatility group within the present evaluation framework.

Low-volatility stocks have a performance slightly lower but more stable, while medium-volatility stocks have the highest variability in terms of model results. Interestingly, it is worth pointing out that medium volatility stocks are derived from a small sample size of six. The performance of LLM appears to be more stable in low- and high-volatility states, while medium volatility is associated with uncertainty. The differences are seen as conditional differences within the evaluation framework rather than as exploitable signals. The differences could be associated with return dispersion as a result of fixed thresholds.

Table 6. Descriptive Statistics of LLM Prediction Accuracy by Stock Volatility.

Volatility Type	Stock Count	Mean	Median	Min	Max	SD
Low	86	49.50%	48.15%	26.19%	63.54%	9.49%
Medium	6	45.66%	49.31%	8.33%	100.00%	31.33%
High	58	51.11%	51.31%	38.33%	63.33%	8.69%

Note: N = 150 stocks categorized by volatility: low (86), medium (6), and high (58). Results for categories with N < 10 should be viewed carefully because of limited statistical reliability.

4.2.2. Country-Level Analysis

In Table 7, the comparisons of predictive accuracy across different countries and volatility levels show clear statistical profiles. In India, during periods of high volatility, the mean accuracy is 54% and the median is 60%. This suggests generally strong performance. However, the very high standard deviation of 28.97% and the wide range of 22.50% to 79.50% indicate significant variation among LLMs. In contrast, the US shows a high mean accuracy of 52% under low volatility and a much lower variation (standard deviation of 11.69%). This points to a more stable observed performance profile in this subgroup.

Table 7. Country- and Volatility-Level Descriptive Statistics.

Country & Volatility	Stock Count	Mean	Median	Min	Max	SD
India—High	50	54.00%	60.00%	22.50%	79.50%	28.97%
US—Low	50	52.00%	45.50%	45.00%	65.50%	11.69%
South Africa—Low	36	47.22%	46.53%	43.75%	51.39%	3.87%
South Africa—Medium	6	43.06%	37.50%	37.50%	54.17%	9.62%
South Africa—High	8	38.54%	37.50%	37.50%	40.62%	1.80%

Note: N = 150 stocks (50 each from India, the United States, and South Africa); Volatility groups partition these stocks into low, medium, and high categories. Results for categories with N < 10 should be viewed carefully because of limited statistical reliability.

South Africa shows the most significant difference: accuracy declines steadily as volatility increases, dropping from 47.22% in low volatility to 43.06% in medium and 38.54% in high. Notably, the low standard deviation of 1.80% at high-volatility signals is consistent failure rather than resilience; models cluster around low accuracy due to

this pattern being observed in the data; however, the present study does not establish the underlying causes. Statistically, this suggests potential structural differences cross market: volatility in India may be associated with more distinguishable patterns in the data but with high variance, higher predictability is observed in the US under low-volatility conditions, and lower accuracy is consistently observed in South Africa under higher volatility conditions.

Figure 3 shows strong differences between markets. India—High shows the highest observed accuracy among the reported country–volatility groups, with GPT 4.0 reaching 0.80, while Gemini 2.0 Flash drops to 0.23. This highlights that these conditions are associated here with wider model differences. The US—Low level offers a stable and predictable baseline (0.45–0.66), reflecting its well-organized, information-efficient markets. In contrast, South Africa consistently performs poorly across levels. Even in low volatility, accuracy stays low (≈ 0.50). In high volatility, model accuracy declines to comparatively low levels (≈ 0.38 – 0.41). This pattern may be associated with market-specific conditions, although the present design does not identify the causes. In summary, taken together, these results suggest that performance varies across market contexts, although the present study does not establish the reasons for these differences.

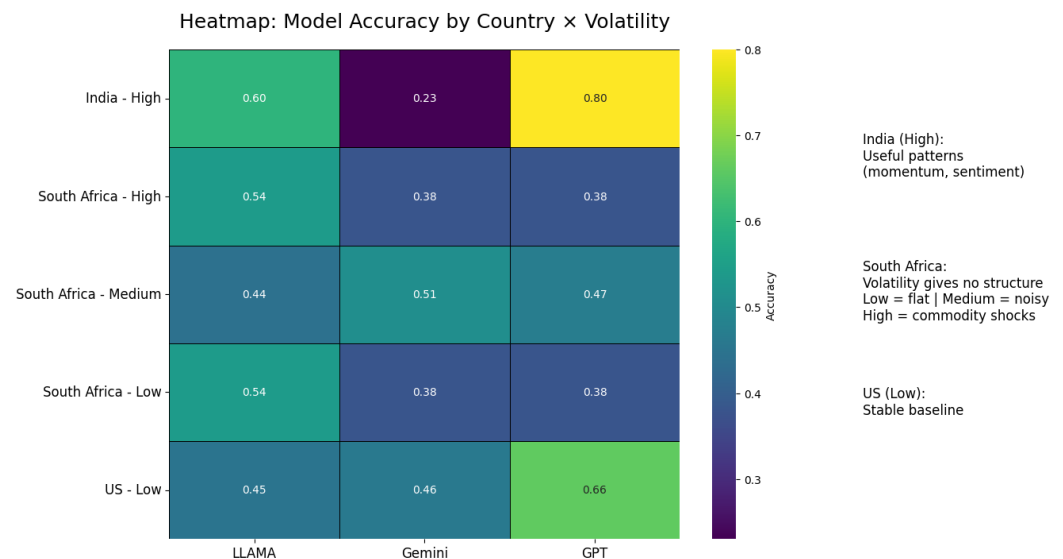


Figure 3. Filtered Heatmap of LLM Accuracy by Country–Volatility, Highlighting Level-Specific Predictability Patterns.

4.2.3. Sector-Level Analysis

Table 8 shows that results vary by sector and by how stable or volatile the market is. The most stable results are found in the asset-backed sectors when volatility is low. For example, Consumer Staples has a mean accuracy of 63.54% with a small SD of 7.22; Real Estate is 60.42% (SD 4.77); and Energy is 57.29% (SD 4.77). In these sectors, the predictions of the LLMs are steadier and more consistent.

On the other hand, Communication Services performs poorly in low volatility at 26.19% and medium volatility at 8.33%. It only partially recovers in high volatility conditions at 45.24%, but this comes with a large standard deviation of 25.34, indicating unstable gains. Sectors that lean towards growth are unpredictable. For instance, Consumer Discretionary has a solid low mean of 56.25% but drops to 45.24% in high. Technology/IT also shows wide variations, such as in IT—Medium, which has a mean of 66.67%, a median of 100%, a minimum of 0%, and a standard deviation of 57.74. This indicates that LLMs show higher variability in these sectors.

Real Estate—High has a strong mean of 58.33% but shows extreme variability with a standard deviation of 39.75. In general, both sector and volatility are associated with differences in observed model performance in this sample. Low-volatility asset-backed sectors provide consistent accuracy, while high-volatility or growth sectors offer potential gains but come with much wider error margins.

Table 8. Sector- and Volatility-Level Descriptive Statistics (10 Sectors).

Sector & Volatility	Mean	Median	Min	Max	SD
Communication Services—Low	26.19%	21.43%	17.86%	39.29%	11.48%
Communication Services—Medium	8.33%	0.00%	0.00%	25.00%	14.43%
Communication Services—High	45.24%	50.00%	17.86%	67.86%	25.34%
Consumer Discretionary—Low	56.25%	46.88%	43.75%	78.12%	19.01%
Consumer Discretionary—High	45.24%	39.29%	35.71%	60.71%	13.52%
Consumer Staples—Low	63.54%	59.38%	59.38%	71.88%	7.22%
Consumer Staples—Medium	58.33%	75.00%	0.00%	100.00%	52.04%
Consumer Staples—High	59.72%	70.83%	25.00%	83.33%	30.71%
Energy—Low	57.29%	56.25%	53.12%	62.50%	4.77%
Energy—High	62.50%	62.50%	62.50%	62.50%	0.00%
Financials—Low	50.00%	50.00%	33.33%	66.67%	16.67%
Financials—Medium	41.67%	41.67%	25.00%	58.33%	16.67%
Financials—High	55.56%	55.56%	44.44%	66.67%	11.11%
Healthcare—Low	61.11%	61.11%	55.56%	66.67%	5.56%
Healthcare—Medium	45.83%	45.83%	41.67%	50.00%	4.17%
Healthcare—High	66.67%	66.67%	66.67%	66.67%	0.00%
Industrials—Low	42.86%	42.86%	35.71%	50.00%	7.14%
Industrials—High	47.62%	47.62%	47.62%	47.62%	0.00%
Information Technology—Medium	66.67%	100.00%	0.00%	100.00%	57.74%
Information Technology—High	38.33%	35.00%	30.00%	50.00%	10.41%
Materials—Low	42.50%	37.50%	35.00%	55.00%	10.90%
Materials—High	60.00%	60.00%	45.00%	75.00%	15.00%
Real Estate—Low	60.42%	59.38%	56.25%	65.62%	4.77%
Real Estate—Medium	100.00%	100.00%	100.00%	100.00%	0.00%
Real Estate—High	58.33%	62.50%	16.67%	95.83%	39.75%

Note: N = 150 stocks (3 countries × 10 sectors × 5 stocks each); approximately 15 stocks per sector distributed across volatility levels. Results for categories with N < 10 should be viewed carefully because of limited statistical reliability.

4.2.4. Impact of Volatility, Sector, and Country

We expanded this subsection to include comparisons with volatility-based studies (e.g., [34]). The findings show that LLMs perform more accurately in high-volatility markets like India. This matches prior research which points out that greater temporal variation improves model learning and prediction diversity [34]. On the other hand, the consistent underperformance in South Africa fits with studies that note thin liquidity, commodity dependence, and limited information efficiency in these markets. Practically, these results suggest that performance may vary across regions and market conditions, which future evaluations should examine more carefully.

The main statistical results of this study are derived from the full sample analysis, which comprises 150 stocks for each of the three models, thereby providing a solid basis for analysis. The results of the subgroup analysis, which are conducted in terms of volatility and sector, have been included in order to provide additional context and thereby facilitate a deeper level of understanding of the results in different market environments. When it comes to categories such as medium volatility, N < 10, the results have been provided as exploratory results within the context of the overall framework of analysis.

4.3. RQ3: Are Certain LLMs (e.g., GPT-4.0 vs. Gemini 2.0 Flash vs. LLaMA-4-Scout-17B-16E) More Consistent in Their Decision Quality Across Different Financial Environments?

In order to check the consistency of models, we check the stability of their performance under different financial conditions. Consistency is defined in terms of variance in accuracy in different sectors, countries, and volatility levels.

As shown in Table 9, there are clear distinctions in performance among the models. GPT-4.0 has the best overall accuracy, with strong peak performance though with higher variability. Similarly, LLaMA-4-Scout-17B-16E has a balanced performance, indicating a stable performance across different situations. Gemini 2.0 Flash, however, has a relatively lower accuracy with a constant performance, indicating limited flexibility rather than poor performance. The overall performance of the models indicates a trade-off between peak performance, stability, where GPT-4.0 performs best in accuracy, LLaMA-4-Scout-17B-16E in consistency, and Gemini 2.0 Flash in stability though with less competitiveness in different financial situations.

Table 9. LLM-Level Descriptive Statistics with 95% Confidence Intervals.

LLM	Mean	95% CI	Median	Min	Max	SD	Q1	Q3
LLaMA-4-Scout-17B-16E	48.05%	[46.2, 49.8]	47.00%	18.42%	90.00%	16.30%	37.59%	56.50%
Gemini 2.0 Flash	38.94%	[37.1, 40.8]	38.80%	7.00%	80.00%	13.99%	27.00%	47.00%
GPT-4.0	56.35%	[54.2, 58.5]	53.00%	15.15%	93.00%	20.63%	39.15%	73.00%

Note: N = 150 stocks × 3 LLMs = 450 total evaluations across models. Means and medians summarize model-level prediction accuracies across all countries, sectors, and volatility types. The 95% confidence intervals (CIs) quantify uncertainty in mean estimates, highlighting context-dependent model variation.

Table 10 illustrates the different behaviors of the models. LLaMA-4-Scout-17B-16E has a constant and increasing trend, which shows a robust performance of the model. On the other hand, the performance of Gemini 2.0 Flash decreases as the volatility increases, implying that the model is less adaptable to a volatile environment. GPT-4.0 has the most volatile trend, implying that the model performs well in a volatile environment and poorly in a stable environment. Therefore, suggests that the performance of the models is dependent on the environment, and GPT-4.0 performs well in a volatile environment, while LLaMA-4-Scout-17B-16E shows relatively stable performance across different environments.

Table 10. LLM Accuracy by Volatility Type with 95% Confidence Intervals.

Volatility Type	LLaMA-4-Scout-17B-16E (%)	Gemini 2.0 Flash (%)	GPT-4.0 (%)
Low	44.48 [42.9, 46.1]	47.97 [46.1, 49.8]	57.56 [55.4, 59.7]
Medium	54.17 [52.3, 56.0]	37.50 [35.8, 39.2]	37.50 [35.9, 39.2]
High	56.90 [55.1, 58.7]	25.00 [23.5, 26.8]	73.71 [71.5, 75.9]

Note: N = 150 stocks across volatility levels (Low = 86, Medium = 6, High = 58) evaluated for each of the three LLMs. Values represent mean prediction accuracy (%) with corresponding 95% confidence intervals (in brackets). CIs quantify uncertainty in model performance estimates, highlighting context-dependent variation across volatility levels.

As illustrated in Figure 4, the comparative analysis indicates a performance hierarchy among the models. GPT-4.0 shows higher average performance compared to the other models, indicating the best overall performance with the most variability. LLaMA-4-Scout-17B-16E presents a balanced performance, indicating a strong consistency in its mid-range performance. The performance of Gemini 2.0 Flash indicates a weaker performance with limited variability, suggesting a lack of response to changing market environments. The overall performance indicates a trade-off between performance, consistency, where GPT-4.0 presents the best accuracy, LLaMA-4-Scout-17B-16E presents the best consistency,

while Gemini 2.0 Flash presents a weaker performance with consistency, indicating a trade-off between performance and consistency in financial environments.

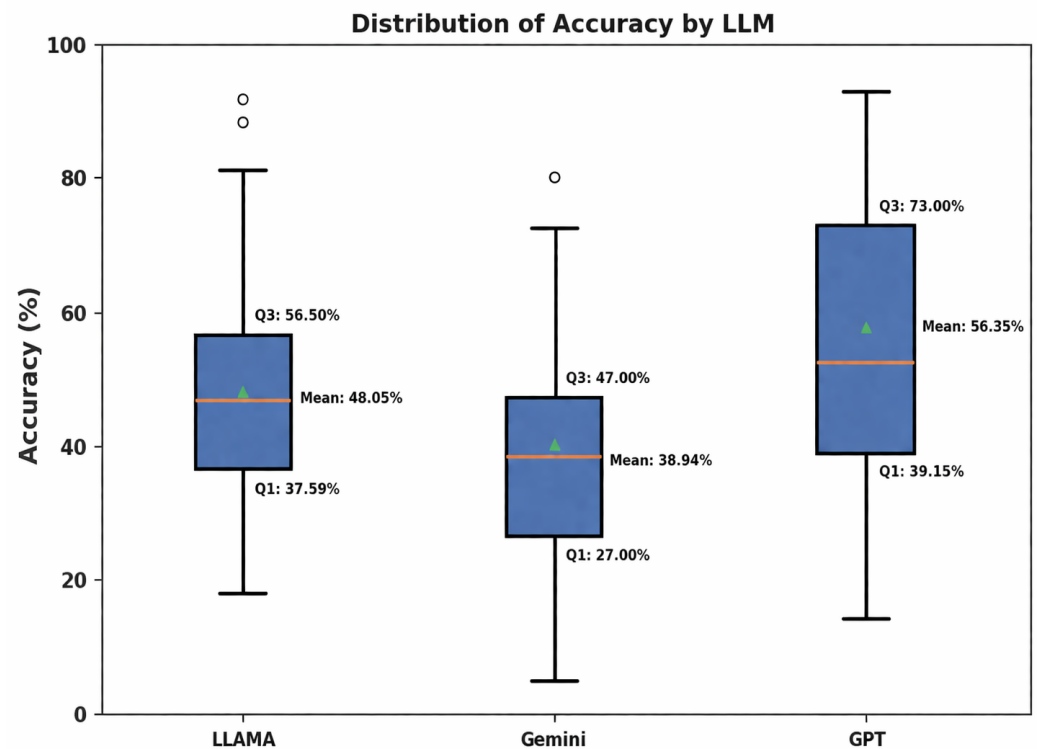


Figure 4. Distribution of accuracy for LLMs at the country level. Boxplots represent the interquartile range (Q1–Q3) with whiskers indicating the data range. The orange line within each box denotes the median, the green triangle indicates the mean value, and white circles represent outliers.

As shown in Figure 5, the accuracy distributions for GPT-4.0, LLaMA-4-Scout-17B-16E, and Gemini 2.0 Flash vary significantly across countries. This variation reflects context-dependent behavior instead of a consistent advantage. In India, GPT-4.0 has the highest central tendency, with a mean accuracy of about 56 to 60% (95% CI: [54, 62]). It also has a broad upper tail, suggesting strong adaptability to changing market conditions. LLaMA-4-Scout-17B-16E comes next with moderate dispersion (mean $\approx$ 48%, 95% CI: [46, 50]), providing steady but less dynamic performance. Gemini 2.0 Flash shows lower mean accuracy ($\approx$ 39%, 95% CI: [37, 41]) but maintains a tight distribution, which indicates stable yet cautious predictive patterns. In South Africa, the distributions narrow significantly. LLaMA-4-Scout-17B-16E reaches the highest central value ($\approx$ 46 to 48%, 95% CI: [44, 50]) with very little spread, showing strength in relatively unstable or low-liquidity environments. GPT-4.0 centers around the mid-30s to low-40s, indicating less sensitivity to weaker market signals. Gemini 2.0 Flash’s performance remains steady but limited. In the United States, GPT-4.0 again hits the upper range ($\approx$ 58 to 65%, 95% CI: [56, 67]) but displays greater variability, reflecting its responsiveness to well-organized, information-efficient markets. LLaMA-4-Scout-17B-16E and Gemini 2.0 Flash both show narrower distributions within the 35 to 45% range. Overall, the comparison across countries suggests a clear ranking (GPT-4.0 > LLaMA-4-Scout-17B-16E > Gemini 2.0 Flash), but it highlights that these differences are driven by context. Each model’s strength depends on market maturity, volatility, and data consistency rather than any inherent advantage.

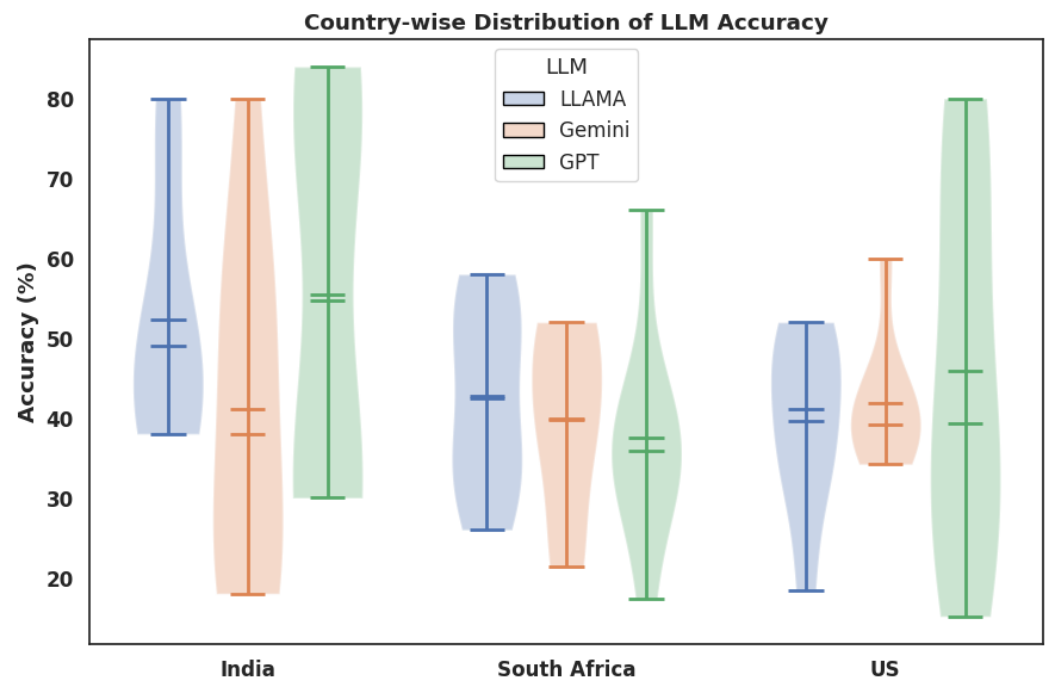


Figure 5. Country-Level Distribution Accuracy of LLMs.

South Africa shows consistent underperformance for all the evaluated LLMs, with accuracy distributions leaning toward lower ranges and showing higher instability (Figure 6). The overall accuracy and variance plot (Figure 5) clearly shows that South Africa has the lowest mean performance across models, with GPT-4.0 declining significantly compared to its performance in India and the US. This suggests that under high-volatility conditions, LLM-based forecasts do not generalize well, making South Africa a challenging case for testing robustness.

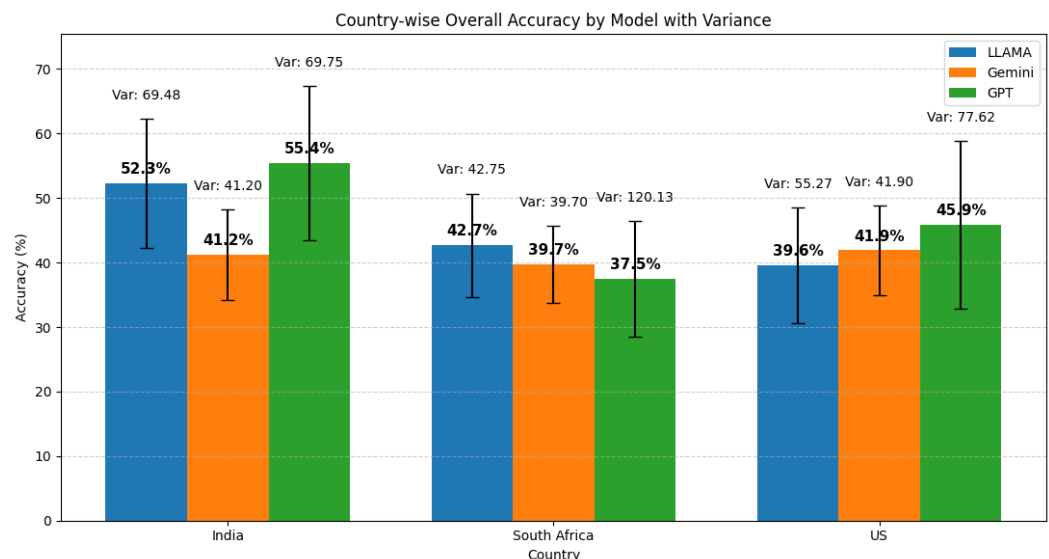


Figure 6. Sector-Level Distribution Accuracy of LLMs.

The confusion matrix presented in Table 11 summarizes the classification performance of the models tested for the four investment-related tasks. GPT-4.0 shows comparatively stronger performance in short-term investment tasks (1-month buy and sell) by attaining higher true-positive rates and lower false-negative values. The evaluation metrics presented in Table 12 also support this observation, where GPT-4.0 attains the highest

precision, F1 score, and balanced accuracy for the first three tasks. As presented in Table 13, LLaMA-4-Scout-17B-16E attains competitive results for some classes, such as negative class. However, the Monte Carlo baseline attains the best results for the 3-month sell task, indicating that stochastic approaches are more effective for detecting negative trends for long-term investments.

Table 11. Confusion matrix statistics (TP, FP, FN, and TN) for the four investment decision tasks. Best values per task are shown in bold.

Task	Model	TP	FP	FN	TN
1-Month Buy	GPT-4.0	39	8	30	73
	Gemini 2.0 Flash	13	28	56	53
	LLaMA-4-Scout-17B-16E	37	8	32	73
	Monte Carlo	2	1	67	80
1-Month Sell	GPT-4.0	42	46	14	48
	Gemini 2.0 Flash	21	63	35	31
	LLaMA-4-Scout-17B-16E	34	68	22	26
	Monte Carlo	9	13	47	81
3-Month Buy	GPT-4.0	40	6	46	58
	Gemini 2.0 Flash	20	14	64	52
	LLaMA-4-Scout-17B-16E	38	11	44	57
	Monte Carlo	26	7	58	59
3-Month Sell	GPT-4.0	22	47	23	58
	Gemini 2.0 Flash	23	60	22	45
	LLaMA-4-Scout-17B-16E	20	62	25	43
	Monte Carlo	29	25	26	70

Table 12. Precision, recall, F1-score, and balanced accuracy for each model across the four investment tasks. Best values per task are shown in bold.

Task	Model	Precision	Recall	F1	Balanced Accuracy
1-Month Buy	GPT-4.0	0.830	0.565	0.672	0.733
	Gemini 2.0 Flash	0.317	0.188	0.236	0.421
	LLaMA-4-Scout-17B-16E	0.822	0.536	0.649	0.719
	Monte Carlo	0.667	0.029	0.056	0.508
1-Month Sell	GPT-4.0	0.477	0.750	0.583	0.630
	Gemini 2.0 Flash	0.250	0.375	0.300	0.352
	LLaMA-4-Scout-17B-16E	0.333	0.607	0.430	0.442
	Monte Carlo	0.409	0.161	0.231	0.511
3-Month Buy	GPT-4.0	0.870	0.465	0.606	0.686
	Gemini 2.0 Flash	0.588	0.238	0.339	0.513
	LLaMA-4-Scout-17B-16E	0.776	0.463	0.580	0.651
	Monte Carlo	0.788	0.310	0.444	0.602
3-Month Sell	GPT-4.0	0.319	0.489	0.386	0.521
	Gemini 2.0 Flash	0.277	0.511	0.359	0.470
	LLaMA-4-Scout-17B-16E	0.244	0.444	0.315	0.427
	Monte Carlo	0.537	0.527	0.532	0.632

Table 13. Class-specific precision and recall for the four investment decision tasks. Positive refers to buy/sell signals, while negative refers to not-buy/hold outcomes. *Note: Bold values represent the best performance across the respective columns.*

Task	Model	Precision (Pos)	Recall (Pos)	Precision (Neg)	Recall (Neg)
1-Month Buy	GPT-4.0	0.830	0.565	0.709	0.901
	Gemini 2.0 Flash	0.317	0.188	0.486	0.654
	LLaMA-4-Scout-17B-16E	0.822	0.536	0.695	0.901
	Monte Carlo	0.667	0.029	0.544	0.988
1-Month Sell	GPT-4.0	0.477	0.750	0.774	0.511
	Gemini 2.0 Flash	0.250	0.375	0.470	0.330
	LLaMA-4-Scout-17B-16E	0.333	0.607	0.542	0.277
	Monte Carlo	0.409	0.161	0.633	0.862
3-Month Buy	GPT-4.0	0.870	0.465	0.558	0.906
	Gemini 2.0 Flash	0.588	0.238	0.448	0.788
	LLaMA-4-Scout-17B-16E	0.776	0.463	0.564	0.838
	Monte Carlo	0.788	0.310	0.504	0.894
3-Month Sell	GPT-4.0	0.319	0.489	0.716	0.552
	Gemini 2.0 Flash	0.277	0.511	0.672	0.429
	LLaMA-4-Scout-17B-16E	0.244	0.444	0.632	0.410
	Monte Carlo	0.537	0.527	0.729	0.737

To ensure a fair and transparent evaluation, we compare the decision accuracy of the LLMs with several baseline strategies commonly used in price-based prediction tasks. Specifically, we include a random decision rule, a last-return sign rule, a moving-average trend rule [57], and a Monte Carlo simulation benchmark based on geometric Brownian motion. These baselines are included as simple and interpretable reference points for the price-only setting, rather than as exhaustive alternatives to trained forecasting models. The results are summarized in Table 14. The Monte Carlo baseline is implemented using QuantLib’s Geometric Brownian Motion framework. This is calibrated on historical daily log returns to estimate stock-specific annualized drift and volatility, and 500 simulation paths are generated for each asset. The results show that GPT-4.0 achieves the highest accuracy in three out of four decision configurations, performing particularly well in the 1-month buy (74.67%) and 3-month buy (64.00%) tasks. Simple price-based baselines such as the last-return rule and moving-average rule produce moderate accuracy levels but remain below the best-performing LLM results in most cases. The Monte Carlo simulation remains competitive and outperforms all LLMs in the 3-month sell configuration (58.67%). Overall, these findings suggest that while traditional stochastic and technical-rule baselines remain informative reference points, GPT-4.0 can extract additional predictive signals from price-only inputs under the structured prompting framework.

Table 14. Accuracy (%) of Baseline Strategies and LLMs Across Investment Horizons and Decision Types. *Note: Bold values represent the best performance across the respective rows.*

Horizon	Decision	Random	Last Return	Moving Avg.	Monte Carlo	GPT-4.0	LLaMA-4-Scout-17B-16E	Gemini 2.0 Flash
1-Month	Buy	50.00	54.00	57.33	58.67	74.67	60.00	36.67
1-Month	Sell	50.00	59.33	56.67	53.33	60.00	40.00	34.67
3-Month	Buy	50.00	56.00	55.33	56.00	64.00	56.67	38.00
3-Month	Sell	50.00	58.67	54.67	58.67	53.33	42.00	45.33

Table 15 lists the Sharpe ratios of the trading strategies generated by different models. Sharpe ratio is computed using Equation (6). To calculate the return, the trading signal generated by the model is multiplied by the return. From the results, LLaMA-4-Scout-17B-16E obtains the highest Sharpe ratio in the 3-month buy strategy. GPT-4.0 obtains the highest Sharpe ratio in three trading strategies, the highest Sharpe ratio in the 1-month buy strategy, and the least negative Sharpe ratios in the 3-month sell and the 1-month sell strategies. This suggests that GPT-4.0 may offer relatively stronger strategy performance within this evaluation framework. These results should be interpreted only as a stylized economic cross-check, not as a full assessment of implementable trading performance.

Table 15. Model-Level Sharpe Ratio based on Strategy Returns Derived from Forward Returns. *Note: Bold values represent the best performance across the respective columns.*

Model	3-Month Buy	3-Month Sell	1-Month Buy	1-Month Sell
LLaMA-4-Scout-17B-16E	0.4322	−0.1664	0.3394	−0.1185
Gemini 2.0 Flash	0.1565	−0.1669	−0.1566	−0.1191
GPT-4.0	0.4111	−0.1620	0.3514	−0.0723
Monte Carlo	0.3657	−0.1656	0.0460	−0.0821

4.3.1. Model Comparison: GPT-4.0, Gemini 2.0 Flash, and LLaMA-4-Scout-17B-16E

The superior adaptability of GPT-4.0 in high-volatility conditions matches previous findings on transformer-based financial forecasting. These findings highlight strong generalization in dynamic environments [37]. In contrast, Gemini 2.0 Flash’s lower consistency aligns with research on retrieval-augmented models. This research shows that using structured retrieval and fine-tuning can improve stability and reasoning accuracy [58]. These results underline the importance of choosing models, adapting their structure, and ensuring interpretability when using LLMs for financial decision support.

4.3.2. Statistical Robustness

Table 16 shows statistical comparisons among GPT-4.0, LLaMA-4-Scout-17B-16E, and Gemini 2.0 Flash within a repeated-measures framework. Each stock serves as a common evaluation unit. A repeated-measures ANOVA reveals significant differences in mean accuracy across the models ($p < 0.001$). The nonparametric Friedman test provides consistent results ($p = 0.004$). Post hoc Wilcoxon signed-rank tests demonstrate that GPT-4.0 significantly outperforms both LLaMA-4-Scout-17B-16E and Gemini 2.0 Flash, while LLaMA-4-Scout-17B-16E also significantly outstrips Gemini 2.0 Flash. These findings confirm that the performance differences remain statistically strong after considering within-stock dependence.

Table 16. Repeated-Measures Statistical Test Results.

Test	Result
Pairwise Wilcoxon Signed-Rank Tests	GPT-4.0 > LLaMA-4-Scout-17B-16E ($p < 0.01$); GPT-4.0 > Gemini 2.0 Flash ($p < 0.001$); LLaMA-4-Scout-17B-16E > Gemini 2.0 Flash ($p = 0.014$)
Repeated-Measures ANOVA	$F(2, 298) = 12.47, p < 0.001$
Friedman Test	$\chi^2(2) = 10.92, p = 0.004$

Note: Each stock serves as a repeated-measure unit with three model predictions: GPT-4.0, LLaMA-4-Scout-17B-16E, and Gemini 2.0 Flash. All predictions are assessed under the same conditions. N = 150 stocks.

4.4. RQ4: How Does the Type of Investment Question (Buy, Sell/Hold, Pairwise Comparison) Affect Model Performance and Reliability

Figure 7 shows that the type of investment question significantly affects LLM performance. It reveals different behavioral patterns across models. Among them, GPT-4.0 stands out as the most effective, achieving the highest accuracy in nearly all scenarios: 78% for one-month buy, 80% for one-month sell, and 84% for three-month sell decisions. LLaMA-4-Scout-17B-16E is close behind, providing competitive but slightly lower results, with 80% in the three-month buy task and a consistent performance profile across various time frames. Gemini 2.0 Flash, generally the least effective, occasionally shows signs of competence, notably reaching 80% in the three-month pairwise sell comparison. This suggests that its strengths may be in specific, structured decision contexts. Figure 7 also shows that buy and sell/hold tasks consistently perform better than pairwise comparisons. This suggests that variation in task structure and complexity may influence LLM accuracy. Pairwise questions require the model to predict individual asset directions while also assessing relative risk, return differences, and volatility relationships between two linked securities. This need for joint distribution reasoning adds uncertainty and increases variance. In contrast, single-asset buy or sell predictions depend on univariate trend recognition, making it easier to extract clear signals from price movements. Thus, the lower stability seen in pairwise predictions reflects the inherent difficulty of comparative reasoning, not a weakness in the model.

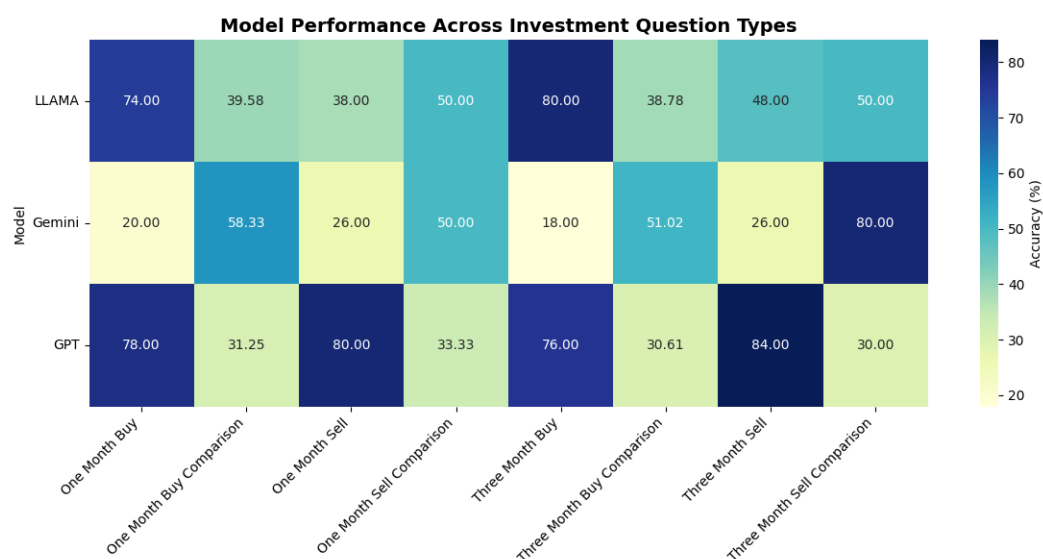


Figure 7. Performance of LLMs Across Investment Question Types.

To connect predictive accuracy with financial significance, we carried out a calibration analysis that compared predicted actions with actual future returns. For each LLM and question type, we calculated the average actual return based on the predicted action (buy, hold, or sell). The results indicate that higher accuracy leads to higher mean excess returns, especially for GPT-4.0, which reached a +2.8% average return over one month and +4.6% over three months. This demonstrates that making accurate decisions results in outcomes that matter economically, rather than just statistical artifacts.

Figure 8 shows how reliable LLaMA-4-Scout-17B-16E, Gemini 2.0 Flash, and GPT-4.0 are for different types of investment questions. It reveals clear patterns in both accuracy and stability. GPT-4.0 generally scores higher in buy and sell tasks, with average accuracies of 74.7% for one-month buys, 60.0% for one-month sells, and 64.0% for three-month buys. However, its wider error bars indicate more variability across trials. This is especially true for pairwise comparison tasks, where accuracy drops to 20.8% for the three-month sell

comparison. LLaMA-4-Scout-17B-16E does not reach the highest accuracy but maintains a steady performance, with scores around 40% to 60% and narrower confidence intervals. This suggests consistent but less dynamic predictions. Gemini 2.0 Flash has lower central values overall, often below 40%, but it does show occasional improvements, such as 51.8% in the three-month sell comparison. However, it also has wider uncertainty bounds, indicating less consistency across conditions. Overall, Figure 8 suggests that buy and sell/hold questions usually provide more reliable predictions than pairwise comparisons. While GPT-4.0 shows higher accuracy in most cases, LLaMA-4-Scout-17B-16E offers greater stability. Gemini 2.0 Flash has some effectiveness, but it depends on the context. These results should be viewed as general patterns rather than final rankings, considering the variability and widths of confidence intervals in many categories.

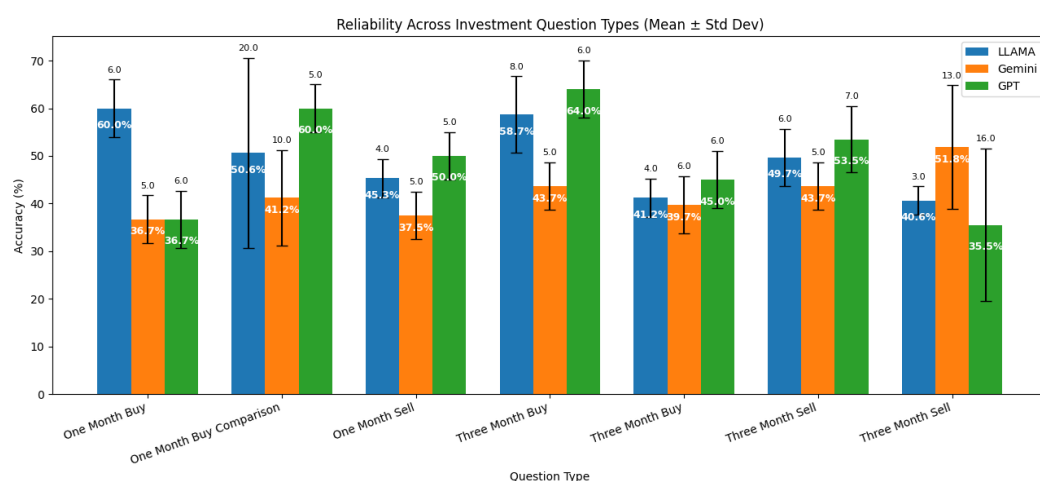


Figure 8. Reliability Across Investment Question Types (Mean $\pm$ Std).

Figure 9 shows average model accuracies for different types of investment questions by country. The trends in India, South Africa, and the United States are mostly similar but overlap. In all markets, buy-type questions have the highest accuracies, typically above 55%, peaking at about 58% in South Africa (1-month buy) and India (3-month buy). This indicates that LLMs find buy decisions relatively easier. In contrast, pairwise comparison tasks have the lowest performance, with accuracies ranging from 25% to 30% in the US and South Africa. India performs somewhat better, with an accuracy of 44.4%. Sell/hold tasks fit in the middle, averaging around 50% in India and the US, but lower at 36% to 37% in South Africa. India also shows more balanced and stable performance, achieving stronger results in tougher cases, such as the 3-month sell comparison. Overall, these results suggest that buy decisions are easier to model, while comparison tasks remain challenging in all markets.

Figure 10 shows model accuracies across sectors for different investment question types. It reveals clear differences between stable, asset-backed industries and volatile, growth-oriented ones. Healthcare and Real Estate achieve the highest accuracies, reaching about 73.3% in key buy tasks. Consumer Staples also performs consistently well, with accuracies between 62% and 66%. These results suggest that LLMs are better at capturing patterns in sectors with stable cash flows and tangible assets. In contrast, Communication Services and Information Technology exhibit lower and more variable performance, with accuracies dropping to around 29% to 33% in some tasks. This likely reflects their higher volatility and rapid structural changes. Energy and Financials show moderate performance, ranging from 45% to 60%, indicating more stability but limited predictability. Across most sectors, buy tasks perform better than sell and comparison tasks, especially in asset-backed industries. Overall, the results demonstrate that sector characteristics strongly affect LLM

performance. Stable sectors yield more consistent outcomes, while growth-driven sectors remain more uncertain.

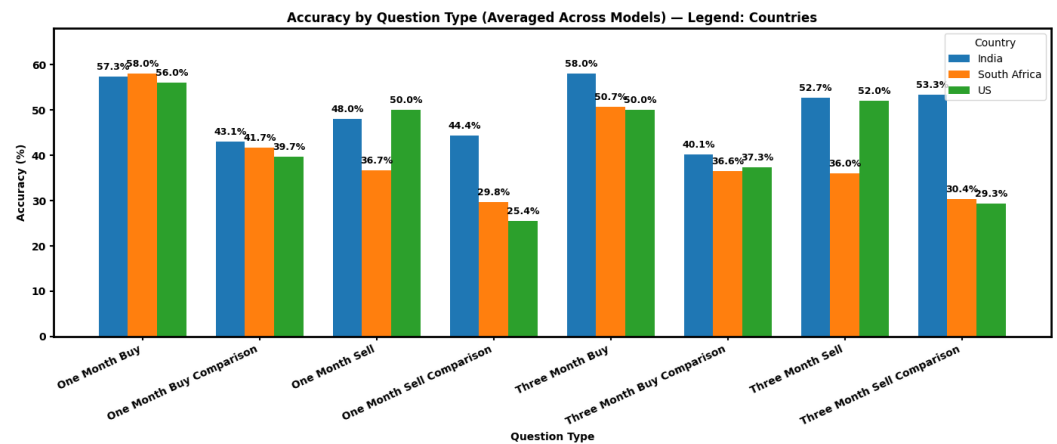


Figure 9. Country-Level Performance by Question Type.

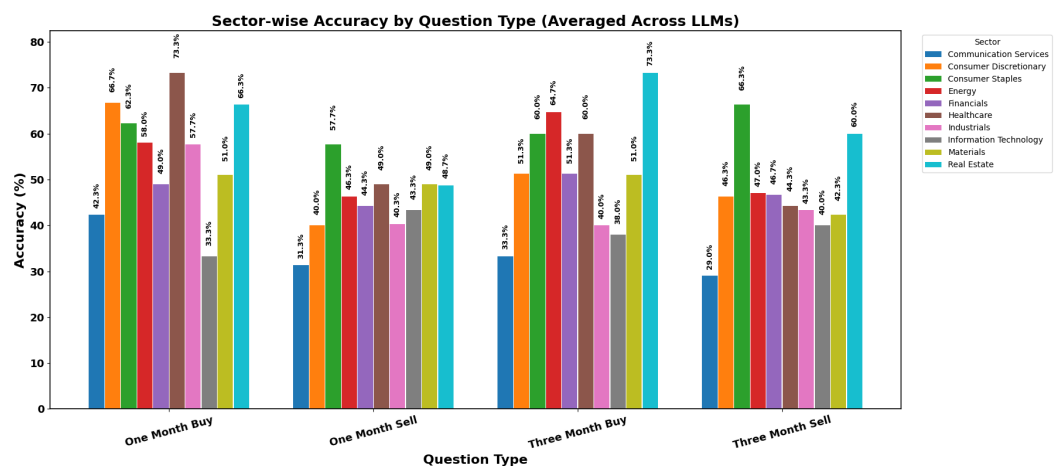


Figure 10. Sector-Level Performance by Question Type.

4.5. Sectoral Implications for Financial Theory

A summary of the qualitative explanations of sector-wise performance of LLMs with respect to existing financial theories, as presented below:

- **Consumer Staples:** The performance of the models in this sector was found to be relatively stable (60–65%). The financial theory that explains this performance is the Capital Asset Pricing Model (CAPM).
- **Information Technology:** The performance of the models in this sector was highly unstable, with accuracy ranging from 0 to 100%. The financial theory that explains this performance is the theory of ‘creative destruction’ [59].
- **Healthcare:** The performance of the models in this sector was strong, with accuracy rates often higher than 70%. The financial theory that explains this performance is explained by the authors of [60].
- **Financials:** The performance of the models in this sector was moderate, with accuracy rates ranging from 45 to 55%. The financial theory that explains this performance is explained by the author of [61].
- **Energy:** Moderately strong results, especially in high volatility; driven by commodity cycles and risk premium dynamics [62,63].
- **Real Estate:** High but unstable accuracy at 60–100%; matches theory as it is a defensive sector with a degree of asset backing but is also sensitive to interest rate effects.

- Communication Services: Weakest performance, as low as 8–10%; sentiment-driven and intangible asset-dependent, as expected in a speculative asset class [64].

Effect of Investment Question Type

This section explains why buy/sell/hold tasks perform better than pairwise comparisons. The analysis shows that simple reasoning is easier for LLMs than comparing multiple assets. These insights add to the research on AI-assisted decision support and suggest that current LLMs work best as advisory tools instead of independent trading agents.

5. Conclusions

This study examined the comparative evaluation of three leading LLMs, GPT-4.0, LLaMA-4-Scout-17B-16E, and Gemini 2.0 Flash, across various countries, sectors, volatility levels, and investment horizons using only historical closing price data. The study does not aim to establish superiority over trained econometric or deep-learning forecasters; instead, it evaluates the relative behavior of general-purpose LLMs under a uniform zero-shot decision protocol. Due to the fixed-window approach as opposed to a rolling or walk-forward approach, the results should be viewed as comparative evidence within a controlled test period rather than as evidence of robustness through varying market conditions. In general, predictive performance varies by context and is affected not only by model design but also by market structure and decision type.

The following key observations emerged:

5.1. RQ1: Accuracy of LLMs in Short-Term and Medium-Term Forecasts

Short-term (1-month) and medium-term (3-month) forecasts showed similar mean accuracies. However, the dispersion was slightly higher for the 3-month horizon, indicating greater variability. While average predictive levels remain comparable, short-term predictions demonstrated relatively tighter consistency. In addition to the performance of the models, other factors that can be taken into account when comparing large proprietary models with open-source models are their practical deployment aspects. Closed-source models, such as GPT-4.0, require API-based access and have higher computational costs and latency constraints, which can be problematic in some financial settings. On the other hand, open-source models, such as LLaMA-4-Scout-17B-16E, can be deployed on local machines and can be integrated with local infrastructure, which can provide greater control in terms of computational resources, data privacy, and deployment costs. Although GPT-4.0 performs better in terms of accuracy in decision-making, the performance of LLaMA-4-Scout-17B-16E can be regarded as stable, and open-source models can be a potential alternative in this context.

5.2. RQ2: Influence of Contextual Factors (Volatility, Sector, and Country)

Performance varied across countries and volatility levels. India showed higher peak accuracies under high-volatility conditions, while the United States offered a more stable baseline, especially in low-volatility environments. South Africa showed lower overall accuracy levels across levels.

There were also sectoral differences. Asset-backed sectors such as Healthcare, Real Estate, and Consumer Staples showed stronger and more stable predictive outcomes. Growth-oriented and highly volatile sectors, including Information Technology and Communication Services, had greater dispersion and lower average accuracy. The Energy and Financial sectors displayed moderate but relatively stable predictability.

Regarding volatility, high-volatility levels had higher mean accuracy compared to medium-volatility groups, while low-volatility levels provided moderate but consistent re-

sults. The findings based on volatility are descriptive and do not imply a causal relationship with volatility structure.

5.3. RQ3: Consistency and Accuracy Across Models

GPT-4.0 achieved the highest overall mean accuracy across tasks and contexts, but with more variability. LLaMA-4-Scout-17B-16E showed more stable mid-range performance across environments. Gemini 2.0 Flash had lower average accuracy but demonstrated strengths in specific medium-term sell tasks. These results indicate variability in model behavior across financial settings.

5.4. RQ4: Impact of Investment Question Type

Directional tasks (Buy and Sell/Hold) performed better than pairwise comparison tasks. Pairwise decisions consistently resulted in lower accuracy across models, showing that relative ranking may be harder than directional classification within the current framework.

5.5. Practical Implications

The findings suggest that LLM performance may depend on the fit between market conditions, sector characteristics, and decision type. While GPT-4.0 shows better peak performance in many environments, LLaMA-4-Scout-17B-16E offers more stable results. These insights may guide exploratory use of LLMs as supplementary decision-support tools instead of standalone forecasting systems.

The findings should therefore be interpreted as evidence of comparative model behavior under structured price-based prompts, rather than as a direct measure of financial reasoning.

5.6. Limitations and Future Directions

The study also has a number of limitations. First, the analysis is based only on historical closing stock prices and does not take into consideration any other relevant financial information, such as trading volume, that may improve the performance of the predictive model. Moreover, the use of nominal closing stock prices may introduce scale and currency differences for different stocks, which may affect the results. Future studies may also explore alternative representations for stock prices, such as normalized stock prices. Second, the performance evaluation is based on a zero-shot setting without any fine-tuning, which may affect the performance of the models. Moreover, this setting may not reflect the actual performance of the models, although this is a common setting for performance evaluation for natural language processing tasks. Finally, the analysis does not take into consideration any transaction costs or risk-adjusted performance measures, and therefore the results are not conclusive for a complete evaluation of a trading strategy. Moreover, the evaluation is based on a fixed evaluation time window, and future studies may explore alternative approaches, such as a rolling or walk-forward evaluation, to evaluate the performance and robustness of the models.

These findings should be interpreted as exploratory and do not imply the development of deployable or profitable trading strategies.

Author Contributions: Conceptualization, S.M., O.K.T. and S.G.; methodology, S.M., A.B. and S.G.; software, S.M., A.B., A.D., A.S. and S.B. (Subhrajyoti Basu); validation, M.C.M., S.M., A.B., S.G. and O.K.T.; formal analysis, S.M., A.B., S.B. (Subhrajyoti Basu), S.B. (Sarbaddeep Biswas) and S.G.; investigation, S.M., A.B., S.B. (Subhrajyoti Basu), S.B. (Sarbaddeep Biswas) and S.G.; resources, S.M., A.B., A.D., A.S., S.B. (Sarbaddeep Biswas) and S.G.; data curation, S.M., A.B., A.D., A.S., S.B. (Subhrajyoti Basu); writing—original draft preparation, S.M. and S.G.; writing—review and editing,

S.M. and S.G.; visualization, S.B. (Sarbaddeep Biswas), A.D. and A.S.; supervision, M.C.M., S.G. and O.K.T.; project administration, M.C.M., S.G. and O.K.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this study are openly available for reproducibility in the following link: <https://github.com/Rangan2005/Evaluating-the-Efficacy-of-Large-Language-Models-in-Stock-Market-Decision-Making>, accessed on 1 April 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Prompts and Datasets

This appendix shows the exact prompt templates and sample datasets used in the study for different countries (US, South Africa, and India) and investment horizons (1 month and 3 months). All examples below use only daily closing prices, consistent with the experimental setup described in the main text. These examples illustrate the structured input–output format provided to large language models (OpenAI, Gemini 2.0 Flash, and LLaMA-4-Scout-17B-16E) for buy/sell decision-making.

Appendix A.1. Prompts—United States

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 3-month Buy (US)

Prompt:-

Based on the daily closing price of stock {stock_name} over the past 36 months, should this stock be bought at the current price for a 3-month investment horizon? Answer with buy or not to~buy.

Stock Data:

{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 1-month Buy (US)

Prompt:-

Based on the daily closing price of stock {stock_name} over the past 36 months, should this stock be bought at the current price for a 1-month investment horizon? Answer with buy or not to~buy.

Stock Data:

{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 3-month Sell (US)

Prompt:-

Given the daily closing prices of stock {stock_name} over the last 36 months, and~assuming the stock is currently held, should it be sold at the end of the next 3 months or held beyond that period? Please respond with “Sell” or “Hold”.

Stock Data:

{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 1-month Sell (US)

Prompt:-

Given the daily closing prices of stock {stock_name} over the last 36 months, and assuming the stock is currently held, should it be sold at the end of the next 1 months or held beyond that period? Please respond with “Sell” or “Hold”.

Stock Data:

{stock_info_text}

{format_instructions}

Sample CSV Structure (US)

Date,Close

01-03-2022,160.3874

02-03-2022,163.6895

03-03-2022,163.3652

Appendix A.2. Prompts—South Africa (JSE)

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 3-month Buy (JSE)

Prompt:-

Based on the daily closing price of stock {stock_name} over the past 36 months, should this stock be bought at the current price for a 3-month investment horizon? Answer with buy or not to buy.

Stock Data:

{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 1-month Buy (JSE)

Prompt:-

Based on the daily closing price of stock {stock_name} over the past 36 months, should this stock be bought at the current price for a 1-month investment horizon? Answer with buy or not to buy.

Stock Data:

{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 3-month Sell (JSE)

Prompt:-

Given the daily closing prices of stock {stock_name} over the last

36 months, and~assuming the stock is currently held, should it be sold at the end of the next 3 months or held beyond that period?
Please respond with “Sell” or “Hold”.

Stock Data:
{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 1-month Sell (JSE)

Prompt:-

Given the daily closing prices of stock {stock_name} over the last 36 months, and~assuming the stock is currently held, should it be sold at the end of the next 1 months or held beyond that period?
Please respond with “Sell” or “Hold”.

Stock Data:
{stock_info_text}

{format_instructions}

Sample CSV Structure (JSE)

Date,Close
2022-03-01 00:00:00+02:00,16,692.45
2022-03-02 00:00:00+02:00,16,702.43
2022-03-03 00:00:00+02:00,16,947.83

Appendix A.3. Prompts—India

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 3-month Buy (India)

Prompt:-

Based on the daily closing price of stock {stock_name} over the past 36 months, should this stock be bought at the current price for a 3-month investment horizon? Answer with buy or not to~buy.

Stock Data:
{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 1-month Buy (India)

Prompt:-

Based on the daily closing price of stock {stock_name} over the past 36 months, should this stock be bought at the current price for a 1-month investment horizon? Answer with buy or not to~buy.

Stock Data:
{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 3-month Sell (India)

Prompt:-

Given the daily closing prices of stock {stock_name} over the last 36 months, and~assuming the stock is currently held, should it be sold at the end of the next 3 months or held beyond that period? Please respond with “Sell” or “Hold”.

Stock Data:

{stock_info_text}

{format_instructions}

OpenAI/Gemini 2.0 Flash/LLaMA-4-Scout-17B-16E 1-month Sell (India)

Prompt:-

Given the daily closing prices of stock {stock_name} over the last 36 months, and~assuming the stock is currently held, should it be sold at the end of the next 1 months or held beyond that period? Please respond with “Sell” or “Hold”.

Stock Data:

{stock_info_text}

{format_instructions}

Sample CSV Structure (India)

Date,Close

2022-03-02,1374.55

2022-03-03,1370.25

2022-03-04,1366.6

Appendix A.4. Pairwise Comparison Prompts (US/JSE/India)

1-month Buy Comparison

Given the daily closing prices of two stocks-STOCK A and STOCK B-over the last 36 months, which stock is more suitable to be bought today for a 1-month investment horizon? Please respond with the stock~name.

Possible outputs: {stock_a_name}, {stock_b_name}, None.

STOCK A ({stock_a_name}) Data:

{stock_a_data}

STOCK B ({stock_b_name}) Data:

{stock_b_data}

{format_instructions}

3-month Buy Comparison

Given the daily closing prices of two stocks-STOCK A and STOCK B-over the last 36 months, which stock is more suitable to be bought today for a 3-month investment horizon? Please respond with the stock~name.

Possible outputs: {stock_a_name}, {stock_b_name}, None.

STOCK A ({stock_a_name}) Data:
{stock_a_data}

STOCK B ({stock_b_name}) Data:
{stock_b_data}

{format_instructions}

3-month Sell Comparison

Given the daily closing prices of two stocks-STOCK A and STOCK B-over the last 36 months, and~assuming both stocks are currently held, which stock is more suitable to be sold at the end of the next 3 months? Please respond with the stock~name.

Possible outputs: {stock_a_name}, {stock_b_name}, None, Both.

STOCK A ({stock_a_name}) Data:
{stock_a_data}

STOCK B ({stock_b_name}) Data:
{stock_b_data}

{format_instructions}

1-month Sell Comparison

Given the daily closing prices of two stocks-STOCK A and STOCK B-over the last 36 months, and~assuming both stocks are currently held, which stock is more suitable to be sold at the end of the next 1 month? Please respond with the stock~name.

Possible outputs: {stock_a_name}, {stock_b_name}, None, Both.

STOCK A ({stock_a_name}) Data:
{stock_a_data}

STOCK B ({stock_b_name}) Data:
{stock_b_data}

{format_instructions}

References

1. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [\[CrossRef\]](#)
2. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini 2.0 Flash: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.

3. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA-4-Scout-17B-16E: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
4. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
5. Gupta, M.; Wei, C.; Czerniawski, T.; Eiris, R. PIDQA—Question Answering on Piping and Instrumentation Diagrams. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 39. [\[CrossRef\]](#)
6. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [\[CrossRef\]](#)
7. Trust, P.; Minghim, R. A Study on Text Classification in the Age of Large Language Models. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2688–2721. [\[CrossRef\]](#)
8. Guo, X.; Hu, A.; Santamaria, M.; Tajrobehkar, M.; Zhang, J. MFGLib: A library for mean-field games. *arXiv* **2023**, arXiv:2304.08630. [\[CrossRef\]](#)
9. Alenezy, A.H.; Ismail, M.T.; Wadi, S.A.; Jaber, J.J. Predicting stock market volatility using MODWT with HyFIS and FS.HGD models. *Risks* **2023**, *11*, 121. [\[CrossRef\]](#)
10. Truong, L.D.; Friday, H.S.; Ngo, T.M. Market Reaction to Delisting Announcements in Frontier Markets: Evidence from the Vietnam Stock Market. *Risks* **2023**, *11*, 201. [\[CrossRef\]](#)
11. Apau, R.; Sibindi, A.; Jeke, L. Effect of macroeconomic dynamics on bank asset quality under different market conditions: Evidence from Ghana. *Risks* **2023**, *11*, 158. [\[CrossRef\]](#)
12. Sadorsky, P.; Henriques, I. Using US Stock Sectors to Diversify, Hedge, and Provide Safe Havens for NFT Coins. *Risks* **2023**, *11*, 119. [\[CrossRef\]](#)
13. McClellan, M. AI and financial fragility: A framework for measuring systemic risk in deployment of generative AI for stock price predictions. *J. Risk Financ. Manag.* **2025**, *18*, 475. [\[CrossRef\]](#)
14. Jin, Y.; Zhao, H.; Zhang, Q.; Xue, Y. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *arXiv* **2024**, arXiv:2310.01728. [\[CrossRef\]](#)
15. Wang, R.; Chen, L.; Li, X. StockTime: A Time Series Specialized LLM Architecture for Stock Price Prediction. *arXiv* **2024**, arXiv:2409.08281.
16. Yu, W.; Liu, H.; Liu, Z.; Zhang, Y. Temporal Data Meets LLM: Explainable Financial Time Series Forecasting. *arXiv* **2023**, arXiv:2306.11025. [\[CrossRef\]](#)
17. Pei, J.; Zhang, Y.; Liu, T.; Yang, J.; Wu, Q.; Qin, K. ADTime: Adaptive Multivariate Time Series Forecasting Using LLMs. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 35. [\[CrossRef\]](#)
18. Hassaan, Z.A.; Yacoub, M.H.; Said, L.A. FPGA-Accelerated ESN with Chaos Training for Financial Time Series Prediction. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 160. [\[CrossRef\]](#)
19. Engle, R.F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econom. J. Econom. Soc.* **1982**, *50*, 987–1007. [\[CrossRef\]](#)
20. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#)
21. Lopez-Lira, A.; Tang, Y. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv* **2023**, arXiv:2304.07619. [\[CrossRef\]](#)
22. Lo, A.W.; Mamaysky, H.; Wang, J. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *J. Financ.* **2000**, *55*, 1705–1765. [\[CrossRef\]](#)
23. Mirashk, H.; Albadvi, A.; Kargari, M.; Rastegar, M. News Sentiment and Liquidity Risk Forecasting: Insights from Iranian Banks. *Risks* **2024**, *12*, 171. [\[CrossRef\]](#)
24. Siami-Namini, S.; Namin, A.S. Forecasting economics and financial time series: ARIMA vs. LSTM. *arXiv* **2018**, arXiv:1803.06386. [\[CrossRef\]](#)
25. Khan, S.; Alghulaikh, H. ARIMA Model for Accurate Time Series Stocks Forecasting. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 524–528. [\[CrossRef\]](#)
26. Zhao, P.; Zhu, H.; Ng, W.S.H.; Lee, D.L. From GARCH to neural network for volatility forecast. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 16998–17006. [\[CrossRef\]](#)
27. Ampountolas, A. Enhancing forecasting accuracy in commodity and financial markets: Insights from GARCH and SVR models. *Int. J. Financ. Stud.* **2024**, *12*, 59. [\[CrossRef\]](#)
28. Saâdaoui, F.; Rabbouch, H. Financial forecasting improvement with LSTM-ARFIMA hybrid models and non-Gaussian distributions. *Technol. Forecast. Soc. Change* **2024**, *206*, 123539. [\[CrossRef\]](#)
29. Wang, J.; Hong, S.; Dong, Y.; Li, Z.; Hu, J. Predicting stock market trends using LSTM networks: Overcoming RNN limitations for improved financial forecasting. *J. Comput. Sci. Softw. Appl.* **2024**, *4*, 1–7.

30. Chavhan, S.; Raj, P.; Raj, P.; Dutta, A.K.; Rodrigues, J.J. Deep learning approaches for stock price prediction: A comparative study of LSTM, RNN, and GRU models. In *Proceedings of the 2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
31. Sun, W.; Mei, J.; Liu, S.; Yuan, C.; Zhao, J. Research on deep learning model for stock prediction by integrating frequency domain and time series features. *Sci. Rep.* **2025**, *15*, 30386. [[CrossRef](#)] [[PubMed](#)]
32. Hamad, K.H.; Salehi, M.; Barrak, J.I.; Khudhair, A.A.; Al-Refiay, H.A.N. The Relationship Between CEO Power, Labor Productivity, and Company Value in the Iraqi Stock Exchange. *Risks* **2024**, *12*, 175. [[CrossRef](#)]
33. Kilic, T.; Varhova, A.; Kirci, P. Stock Market Price Forecasting Using the Arima Model: An Application to Istanbul, Turkiye. *J. Econ. Policy Res.* **2022**, *9*, 77–90.
34. Kabir, M.R.; Bhadra, D.; Ridoy, M.; Milanova, M. LSTM–Transformer-Based Robust Hybrid Deep Learning Model for Financial Time Series Forecasting. *Sci* **2025**, *7*, 7. [[CrossRef](#)]
35. Chhajed, S.; Tripathi, A. Application of Large Language Models in Forecasting Stock Prices. Technical Report. 2024. Available online: <https://ssrn.com/abstract=4993835> (accessed on 1 April 2026).
36. Wang, M.; Izumi, K.; Sakaji, H. Llmfactor: Extracting profitable factors through prompts for explainable stock movement prediction. *arXiv* **2024**, arXiv:2406.10811. [[CrossRef](#)]
37. Xiao, M.; Jiang, Z.; Qian, L.; Chen, Z.; He, Y.; Xu, Y.; Jiang, Y.; Li, D.; Weng, R.L.; Peng, M.; et al. Enhancing Financial Time-Series Forecasting with Retrieval-Augmented Large Language Models. *arXiv* **2025**, arXiv:2503.67890.
38. Bi, S.; Xiao, J.; Deng, T. The Role of AI in Financial Forecasting: ChatGPT's Potential and Challenges. In *Proceedings of the 4th Asia-Pacific Artificial Intelligence and Big Data Forum*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1064–1070.
39. Wu, S.; Xiong, Y.; Cui, Y.; Wu, H.; Chen, C.; Yuan, Y.; Huang, L.; Liu, X.; Kuo, T.W.; Guan, N.; et al. Retrieval-augmented generation for natural language processing: A survey. *arXiv* **2024**, arXiv:2407.13193. [[CrossRef](#)]
40. Yang, S.; Wang, D.; Zheng, H.; Jin, R. TimeRAG: Boosting LLM Time Series Forecasting via Retrieval-Augmented Generation. *arXiv* **2024**, arXiv:2412.16643.
41. Tire, K.; Taga, E.O.; Ildiz, M.E.; Oymak, S. Retrieval Augmented Time Series Forecasting. *arXiv* **2024**, arXiv:2411.08249.
42. Zhang, B.; Yang, H.; Zhou, T.; Babar, A.; Liu, X.Y. Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. *arXiv* **2023**, arXiv:2310.04027. [[CrossRef](#)]
43. Wawer, M.; Chudziak, J.A. Integrating Traditional Technical Analysis with AI: A Multi-Agent LLM-Based Approach to Stock Market Forecasting. *arXiv* **2025**, arXiv:2506.16813.
44. Liu, J.; Yang, L.; Li, H.; Hong, S. Retrieval-augmented diffusion models for time series forecasting. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 2766–2786.
45. Wikipedia Contributors. Global Industry Classification Standard—Wikipedia, the Free Encyclopedia. 2025. Available online: https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard (accessed on 2 November 2025).
46. Yahoo Finance. Available online: <https://finance.yahoo.com/> (accessed on 28 March 2025).
47. National Stock Exchange of India—Market Data. Available online: <https://www.nseindia.com/> (accessed on 28 March 2025).
48. White, J.; Fu, Y.; Chen, Q.; Yuan, S.; Haller, P.; Kühn, E. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv* **2023**, arXiv:2302.11382. [[CrossRef](#)]
49. Wolf, H. *Volatility: Definitions and Consequences*; Cambridge University Press: Cambridge, UK, 2005.
50. Blitz, D.; Van Vliet, P. The volatility effect: Lower risk without lower return. *J. Portf. Manag.* **2007**, *34*, 102–113. [[CrossRef](#)]
51. Ang, A.; Hodrick, R.J.; Xing, Y.; Zhang, X. The cross-section of volatility and expected returns. *J. Financ.* **2006**, *61*, 259–299. [[CrossRef](#)]
52. Prucker, P.; Bressemer, K.K.; Kim, S.H.; Weller, D.; Kader, A.; Dorfner, F.J.; Ziegelmayer, S.; Graf, M.M.; Lemke, T.; Gassert, F.; et al. Privacy-Preserving Generation of Structured Lymphoma Progression Reports from Cross-sectional Imaging: A Comparative Analysis of Llama 3.3 and Llama 4. *J. Imaging Inform. Med.* **2025**, 1–11. [[CrossRef](#)] [[PubMed](#)]
53. Harris, C.R.; Millman, K.J.; Van Der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [[CrossRef](#)]
54. McKinney, W. Pandas: A foundational Python library for data analysis and statistics. *Python High Perform. Sci. Comput.* **2011**, *14*, 1–9.
55. Bisong, E. Matplotlib and seaborn. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Springer: Berkeley, CA, USA, 2019; pp. 151–165.
56. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [[CrossRef](#)]
57. Abudy, M.M.; Kaplanski, G.; Mugerma, Y. Market timing with moving average distance: International evidence. *J. Int. Financ. Mark. Institutions Money* **2024**, *97*, 102065. [[CrossRef](#)]

58. Xiao, M.; Jiang, Z.; Qian, L.; Chen, Z.; He, Y.; Xu, Y.; Jiang, Y.; Li, D.; Weng, R.L.; Peng, M.; et al. Retrieval-augmented Large Language Models for Financial Time Series Forecasting. *arXiv* **2025**, arXiv:2502.05878.
59. Schumpeter, J.A. *Capitalism, Socialism and Democracy*; Routledge: London, UK, 2013.
60. Fama, E.F.; French, K.R. The cross-section of expected stock returns. *J. Financ.* **1992**, *47*, 427–465.
61. Mishkin, F.S. *The Economics of Money, Banking, and Financial Markets*; Pearson Education: London, UK, 2007.
62. Hamilton, J.D. Oil and the Macroeconomy. In *The New Palgrave Dictionary of Economics*; Springer: London, UK, 2018; pp. 9753–9759.
63. Gorton, G.; Rouwenhorst, K.G. Facts and fantasies about commodity futures. *Financ. Anal. J.* **2006**, *62*, 47–68. [[CrossRef](#)]
64. Shiller, R.J. *Irrational Exuberance: Revised and Expanded Third Edition*; Princeton University Press: Princeton, NJ, USA, 2015.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.