Al-Kharaj: Journal of Islamic Economic and Business
Volume 8 (3), 2026: xxx - xxx
P-ISSN: 2686-262X; E-ISSN: 2685-9300

Analysis of Large Language Model-Based Stock Investment Recommendations for LQ45 Stocks

Dian Anggraeni, Muryani Arsal, Muchriana Muchran
Master of Management Postgraduate Program, Muhammadiyah University of Makassar
Email: dhiananggraeny12@gmail.com muryani@unismuh.ac.id muchranmuchriana@gmail.com

Keywords: *Large Language Models; stock investment; investment recommendation; LQ45 Index; artificial intelligence; decision support system.*

Abstract

This study aims to examine the capability of Large Language Models (LLMs) to generate stock investment recommendations in the form of buy, hold, and sell for stocks included in the LQ45 Index of the Indonesia Stock Exchange. A quantitative experimental approach was employed using three generative artificial intelligence models: ChatGPT, Gemini, and DeepSeek. Each model received an identical prompt incorporating stock price information, trading volume, candlestick patterns, and macroeconomic and microeconomic sentiment. The research sample comprised 45 LQ45 constituent stocks observed over 20 trading days in December 2025, resulting in 900 observations for each model and 2,700 observations in total. Model performance was evaluated based on three primary dimensions: recommendation accuracy relative to actual stock price movements, investment returns, and downside risk. The findings indicate discrepancies between LLM-generated recommendations and actual stock price movements, suggesting that none of the three models can predict stock price direction with complete accuracy. Nevertheless, all models generated positive cumulative returns. DeepSeek achieved the highest Compounded Cumulative Return (CCR) at 24.01%, followed by Gemini at 23.58% and ChatGPT at 1.90%. In terms of risk, ChatGPT demonstrated the most favorable performance, recording the lowest total risk at 16.35%, compared with DeepSeek at 22.34% and Gemini at 24.85%. The loss rates were also below 50% across all models, namely 34.11% for DeepSeek, 37.88% for Gemini, and 27.33% for ChatGPT. These findings suggest that LLMs can function as Decision Support Systems for investment decision-making by generating recommendations with the potential to produce positive returns while maintaining relatively moderate risk. However, LLM-generated recommendations should not be interpreted as definitive market forecasts because their performance remains sensitive to market volatility, information dynamics, and rapidly changing financial conditions.

INTRODUCTION

Artificial intelligence (AI) has emerged as one of the most significant disruptive forces in the global financial industry, fundamentally reshaping how financial information is processed and investment decisions are formulated. The emergence of Large Language Models (LLMs), including ChatGPT developed by OpenAI, Gemini developed by Google DeepMind, and DeepSeek developed by High-Flyer, has accelerated the transition from conventional human-centered investment analysis toward human–machine collaboration. Unlike traditional analytical systems that typically address specific and narrowly defined tasks, LLMs can process and synthesize large volumes of textual information, interpret numerical data, identify patterns, and extract market sentiment within a unified analytical environment. These capabilities position LLMs as potentially powerful Decision Support Systems (DSS), namely computer-based systems that integrate data, analytical models, and user interfaces to support human decision-making in complex environments rather than replace human judgment (Solares et al., 2022).

The distinctive value of LLMs in financial decision-making lies in their ability to integrate heterogeneous information sources. Conventional machine learning models are generally developed for specific predictive tasks and require structured datasets and predefined variables. By contrast, LLMs are generative AI models trained on extensive corpora and designed to capture

complex semantic relationships, perform contextual reasoning, and synthesize information from multiple sources. In stock investment settings, these capabilities allow LLMs to process numerical information, such as stock prices and trading volume, together with textual information, including economic news, corporate disclosures, financial reports, and market sentiment. Consequently, LLMs can potentially transform heterogeneous financial information into actionable investment recommendations, such as buy, hold, or sell, while simultaneously providing natural-language explanations for the recommendations. This explanatory capability is particularly relevant to the DSS perspective because it enables investors to understand the rationale underlying an AI-generated recommendation rather than merely receiving an unexplained predictive output.

Nevertheless, the analytical capabilities of LLMs should not be interpreted as evidence of deterministic forecasting ability. LLMs are fundamentally probabilistic models, and their outputs are therefore subject to uncertainty, contextual sensitivity, and variation in response generation. An investment recommendation generated by an LLM may consequently differ from the actual subsequent movement of a stock price. This limitation makes empirical evaluation essential. In addition to examining recommendation accuracy, investment performance should be assessed through realized returns and downside risk. The confidence level associated with each recommendation can also provide supplementary information regarding the strength of the model's internal assessment, although confidence should not automatically be interpreted as equivalent to statistical probability or forecast reliability.

The rapid expansion of AI applications in financial services further underscores the relevance of this research. According to 6Wresearch (2025), the global AI market in financial technology (fintech) reached approximately USD 14.3 billion in 2024 and is projected to expand at a compound annual growth rate (CAGR) of 17.87% to approximately USD 47.5 billion by 2031. This expansion is associated with increasing demand for automated investment research, algorithmic trading, financial analytics, and personalized financial services. Similarly, Gartner (2023) projected that more than 80% of enterprises worldwide would adopt generative AI applications by 2026. These developments indicate that AI is progressively moving from an experimental technology toward an operational component of financial decision-making.

The relevance of AI-assisted investment decision-making is particularly pronounced in Indonesia because of the rapid expansion of the domestic investor base. Data from the Indonesia Central Securities Depository (Kustodian Sentral Efek Indonesia—KSEI, 2025) indicate that the number of capital market investors reached 20.32 million Single Investor Identification (SID) accounts at the end of December 2025, representing a 37% year-on-year increase from 14.87 million SID accounts in 2024. Equity investors accounted for approximately 8.59 million SID accounts, representing an annual increase of approximately 35%. Furthermore, 52.59% of new investors were below 30 years of age, while 57.29% reported monthly incomes between IDR 10 million and IDR 100 million (KSEI, 2025). These demographic characteristics suggest that a substantial proportion of Indonesian investors are relatively young and technologically familiar, making conversational AI tools such as ChatGPT, Gemini, and DeepSeek increasingly accessible as potential sources of investment information and decision support.

However, the rapid expansion of the investor population has not necessarily been accompanied by a comparable improvement in financial literacy. The 2025 National Survey of Financial Literacy and Inclusion conducted by the Financial Services Authority (Otoritas Jasa Keuangan—OJK) in collaboration with Statistics Indonesia (Badan Pusat Statistik—BPS) reported that capital market literacy in Indonesia stood at 17.78%, whereas capital market inclusion was substantially lower at 1.34% (OJK, 2025). This disparity highlights a critical challenge: a growing number of individuals are participating in financial markets while possessing limited knowledge of the analytical processes required to evaluate investment opportunities. Investors with insufficient knowledge may experience difficulties interpreting technical indicators, candlestick formations, trading volume, financial fundamentals, and macroeconomic

developments. Under these circumstances, accessible AI-based DSS tools may offer potential benefits by reducing the informational and analytical barriers faced by retail investors.

The situation becomes more consequential given the substantial participation of retail investors in Indonesia's stock market. Retail investors account for a significant proportion of daily trading activity, yet many do not have access to institutional-grade analytical platforms, dedicated research teams, or costly real-time information systems. Consequently, they may be particularly vulnerable to speculative decision-making and behavioral biases. Prospect Theory provides an important theoretical perspective for understanding such behavior. Kahneman and Tversky (1979) demonstrated that individuals do not evaluate gains and losses symmetrically and tend to exhibit loss aversion, whereby losses exert a stronger psychological impact than equivalent gains. Investors may also exhibit overconfidence, relying excessively on personal intuition or limited information when making investment decisions. The interaction between limited financial literacy, information overload, and behavioral biases may therefore increase the vulnerability of retail investors to suboptimal investment decisions (Chen et al., 2024).

Conversational AI potentially addresses some of these limitations by translating complex financial information into accessible investment signals. An investor can provide a structured query and receive a buy, hold, or sell recommendation accompanied by a natural-language explanation. From the perspective of Bounded Rationality, originally developed by Simon (1957), such systems may help compensate for human cognitive limitations by reducing the amount of information that investors must independently process. Rather than replacing human judgment, LLM-based DSS can potentially function as an analytical intermediary that summarizes complex technical and economic information and presents it in a comprehensible form (Ko et al., 2024). This capability is particularly relevant for retail investors who face constraints in time, financial resources, analytical expertise, and access to professional investment research.

Despite these potential advantages, a fundamental empirical question remains: How reliable are LLM-generated investment recommendations? Specifically, to what extent does a buy recommendation correspond to a subsequent increase in stock prices? What level of investment return could be achieved if investors systematically followed the recommendations? More importantly, what level of downside risk would investors face? These questions are increasingly important as investors can access multiple competing LLM platforms, including ChatGPT, Gemini, and DeepSeek, without necessarily knowing whether the models differ systematically in the quality and reliability of their investment recommendations.

The existing literature provides important but incomplete evidence. Pelster and Val (2024) reported that LLM-generated signals were associated with short-term abnormal returns in the United States stock market, particularly around specific events such as corporate earnings announcements. However, their analysis did not directly evaluate daily actionable buy, hold, and sell recommendations across individual stocks. Ko et al. (2024) demonstrated that ChatGPT-assisted investment strategies could improve portfolio diversification and potentially compete with benchmark performance. Nevertheless, their analysis primarily focused on aggregate portfolio allocation rather than daily stock-level recommendation accuracy. Chen et al. (2023) found that ChatGPT could provide significant information regarding the direction of daily returns, but their experimental design focused on a single AI model and therefore did not establish whether different LLM architectures generate systematically different investment outcomes. Similarly, Rao and Patel (2023) reported that GPT-based models could capture relatively stable sentiment signals from social media, but their framework did not simultaneously incorporate technical indicators such as candlestick patterns and trading volume.

The broader literature on AI-based stock prediction further demonstrates that sophisticated computational models can capture potentially useful market signals. Kumbure et al. (2022), for example, showed that deep learning approaches, including Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), can provide competitive stock price forecasting performance relative to conventional approaches. However, these models differ fundamentally

from publicly accessible conversational LLMs. Deep learning forecasting models are generally bespoke quantitative systems specifically trained on structured historical data, whereas conversational LLMs are general-purpose models that can be directly accessed and queried by ordinary investors. The distinction is important because the practical question facing retail investors is not merely whether a specialized predictive model can forecast stock prices, but whether publicly available generative AI systems can provide useful and economically meaningful investment decision support.

More recent research reinforces the need for systematic comparative evaluation. Chawla et al. (2025) examined the credibility of several leading AI models, including GPT-4, Gemini, and DeepSeek, in assessing investment risk profiles and reported substantial variation across models in their risk assessments. The findings suggest that different AI systems may interpret identical investment contexts differently. Such variation raises the possibility that differences across LLMs may extend beyond risk assessment to the generation of actionable stock recommendations. If models differ systematically in their interpretation of investment risk, they may also differ in their ability to distinguish between buy, hold, and sell opportunities.

Martinez et al. (2026) provide further evidence of this variation by evaluating GPT-4.0, Gemini 2.0 Flash, and LLaMA-4 in generating buy, sell, and hold recommendations using historical price information across multiple countries. Although GPT-4.0 achieved the highest classification accuracy, its accuracy was approximately 56%, while Gemini 2.0 Flash achieved approximately 39%. These results indicate that even relatively advanced LLMs may have limited predictive reliability when historical price information is used as the primary input. Schneider and Yilmaz (2025) similarly documented substantial variability in LLM-generated stock recommendations. Their findings suggest that even identical prompts may produce considerably different recommendations across sessions, while portfolios based on LLM recommendations may underperform conventional market benchmarks on a risk-adjusted basis. Collectively, these findings indicate that the usefulness of LLMs in investment decision-making cannot be inferred from their technological sophistication alone; it must be established through systematic empirical testing.

Against this background, four major research gaps can be identified.

First, there is limited evidence on the daily accuracy of LLM-generated buy, hold, and sell recommendations when technical and sentiment information are simultaneously incorporated into a standardized prompt. Existing studies have tended to rely primarily on historical prices or examine textual information, news, and financial reports separately. For example, Martinez et al. (2026) primarily employed historical price information, whereas Chen et al. (2023) and Rao and Patel (2023) emphasized textual or sentiment-based information. In actual investment decision-making, however, investors typically consider technical market information and economic or corporate news simultaneously. An integrated framework is therefore necessary to evaluate whether LLMs can effectively synthesize heterogeneous information sources.

Second, direct comparative evidence involving multiple leading LLMs remains limited. Although previous studies have examined individual models or compared selected pairs of models, systematic experimentation involving ChatGPT, Gemini, and DeepSeek under identical conditions remains scarce. This gap is particularly important because the models differ in their underlying architectures, training approaches, reasoning mechanisms, and multimodal capabilities. Such differences may produce heterogeneous outputs even when identical financial information and prompts are provided.

Third, empirical evidence from emerging capital markets, particularly Indonesia, remains insufficient. Much of the existing literature focuses on developed markets such as the United States and European markets, with relatively limited evidence from Southeast Asian capital markets. Indonesia provides an important research context because its stock market exhibits distinct institutional, informational, and behavioral characteristics. The volatility of the Indonesian market and its heterogeneous investor composition may create an environment in which AI-based

decision support produces both opportunities and substantial risks. Moreover, evidence concerning market efficiency suggests that information may not always be incorporated into prices instantaneously, creating a relevant context for investigating whether AI systems can identify potentially exploitable signals.

Fourth, the literature lacks sufficiently standardized and fully documented prompt protocols for evaluating LLM-based investment recommendations. The sensitivity of generative AI outputs to prompt formulation creates an important methodological challenge. As demonstrated by Schneider and Yilmaz (2025), seemingly minor changes in prompt formulation can substantially affect the resulting recommendations. Without a standardized prompt protocol, differences between models may reflect variations in prompt design rather than genuine differences in analytical capability. Establishing a transparent and reproducible prompting framework is therefore essential for improving the methodological rigor and replicability of LLM-based financial research.

These gaps have important theoretical and practical implications. From a practical perspective, the increasing use of AI by retail investors creates a risk that technological accessibility may be mistaken for analytical reliability. Investors may incur losses by following inaccurate recommendations, while excessive skepticism may cause them to overlook potentially useful AI-based decision support. From a regulatory perspective, the rapid diffusion of generative AI in investment activities also creates a need for empirical evidence concerning the reliability, limitations, and potential risks of AI-generated financial recommendations. Regulatory institutions such as the Financial Services Authority and the Indonesia Stock Exchange require evidence-based insights to develop appropriate investor protection and AI governance frameworks. From an academic perspective, systematic evidence concerning LLM performance in emerging-market investment environments remains necessary to advance the growing literature on AI in finance.

Accordingly, this study is designed to address the identified gaps through a quantitative experimental framework. Three widely accessible LLMs—ChatGPT, Gemini, and DeepSeek—are evaluated under an identical and fully documented prompting protocol to generate daily buy, hold, or sell recommendations. The standardized prompt incorporates technical information, including closing prices, trading volume, and candlestick patterns, as well as macroeconomic and microeconomic sentiment information. The recommendations are subsequently evaluated through a historical backtesting framework using stocks included in the LQ45 Index of the Indonesia Stock Exchange during December 2025.

The evaluation framework is multidimensional. First, predictive accuracy is assessed by comparing model recommendations with subsequent actual stock price movements using classification-based metrics. Second, investment performance is evaluated through cumulative returns to determine whether recommendations translate into economically meaningful outcomes. Third, investment risk is examined using loss rate, maximum drawdown, and aggregate risk measures. Finally, comparative analysis is employed to identify differences in performance among ChatGPT, Gemini, and DeepSeek. This approach extends previous studies by moving beyond statistical association or isolated prediction accuracy toward an integrated assessment of predictive validity, economic performance, and investment risk.

The study offers contributions at three levels. Theoretically, it extends the emerging literature on LLMs in finance by conceptualizing generative AI as a DSS that integrates technical market information and economic sentiment within an emerging-market investment environment. The study also connects LLM-based decision support with the perspectives of Bounded Rationality and Prospect Theory, thereby providing a broader behavioral and technological framework for understanding AI-assisted investment decisions.

Methodologically, the study contributes a standardized and reproducible prompt protocol combined with a multidimensional evaluation framework encompassing predictive accuracy, realized investment returns, and downside risk. This design enables more meaningful cross-model comparisons and provides a basis for replication in future research across different markets, asset classes, and AI systems.

Practically, the findings provide evidence for retail investors, researchers, financial institutions, and regulators regarding the appropriate role of LLMs in investment decision-making. Rather than treating AI-generated recommendations as autonomous forecasts, the study positions LLMs as potential decision-support instruments whose usefulness must be evaluated in terms of accuracy, profitability, and risk. This perspective is particularly relevant in an environment where generative AI is becoming increasingly accessible and increasingly influential in individual financial decision-making.

Overall, this study addresses a timely research problem at the intersection of generative artificial intelligence, behavioral finance, and investment decision-making. By systematically comparing ChatGPT, Gemini, and DeepSeek using standardized prompts and integrated technical and sentiment information in the Indonesian LQ45 market, the study seeks to determine not merely whether LLMs can generate plausible investment recommendations, but whether those recommendations possess measurable predictive and economic value under actual market conditions.

METHODS

2.1 Research Design

This study employs a quantitative, comparative, and controlled experimental design to evaluate the ability of Large Language Models (LLMs)—ChatGPT, Gemini, and DeepSeek—to generate actionable stock investment recommendations in the form of BUY, HOLD, and SELL signals. The quantitative approach is reflected in the use of measurable performance indicators, including recommendation accuracy, hit ratio, investment return, loss rate, and investment risk. A controlled comparative experiment was conducted by providing all three LLMs with the same standardized prompt, equivalent market information, and identical observation periods. The recommendations generated by each model were subsequently compared with actual stock-price movements on the following trading day. This backtesting framework enabled the study to assess predictive performance, economic returns, and investment risk while minimizing differences attributable to input information or prompt formulation. The unit of analysis was the stock-day observation. Forty-five LQ45 stocks were observed over 20 trading days, producing 900 observations for each LLM and 2,700 AI-generated recommendation observations across the three models.

2.2 Research Setting, Population, and Sample

The study was conducted using stocks listed on the Indonesia Stock Exchange (IDX), with the LQ45 Index serving as the research universe. The observation period covered 20 trading days in December 2025. LQ45 was selected because it consists of relatively liquid and representative stocks, thereby providing a standardized investment universe for evaluating AI-generated recommendations. The population comprised all LQ45 constituent stocks during the observation period. Purposive sampling was employed, with stocks selected based on predetermined criteria, particularly the absence of trading suspension during the observation period. Stocks that experienced suspension were excluded because their subsequent price movements could not be evaluated under normal market conditions. The final sample consisted of 45 stocks, resulting in 900 stock-day observations per LLM and 2,700 observations across ChatGPT, Gemini, and DeepSeek.

Table 3.1. General Characteristics of the Research Sample

Characteristic	Number
Number of stocks	45
Observation period	December 2025

Trading days	20
Observations per LLM	900
Number of LLMs	3
Total AI recommendation observations	2,700

Source: Authors' compilation based on secondary data, 2025.

The sample represented multiple industry sectors, providing sectoral heterogeneity within the LQ45 universe.

Table 3.2. Sectoral Distribution of the Research Sample

Industry Sector	Number of Stocks
Energy	11
Basic Materials	9
Consumer Non-Cyclicals	6
Financials	5
Infrastructure	4
Consumer Cyclicals	3
Industrials	2
Technology	2
Healthcare	2
Properties & Real Estate	1
Total	45

Source: Authors' compilation based on secondary data, 2025.

2.3 Data Sources and Collection Procedures

The study employed secondary financial-market data and experimental LLM outputs. Documentary data included daily opening, high, low, and closing prices, trading volume, historical price movements, candlestick patterns, and relevant macroeconomic and firm-specific information. These data were organized for each stock-day observation and incorporated into a standardized prompt. Experimental data consisted of investment recommendations generated by ChatGPT, Gemini, and DeepSeek. Each model received equivalent information and was instructed to provide a BUY, HOLD, or SELL recommendation, together with a confidence level and supporting rationale. To ensure comparability, each experiment was conducted in a new conversation session, and the same prompt structure was maintained throughout the observation period. All generated recommendations were recorded immediately after generation.

2.4 Standardized LLM Experimental Protocol

A standardized experimental protocol was applied to all three LLMs. The prompt incorporated four principal information categories: (1) stock-price information, (2) trading volume, (3) technical indicators and candlestick patterns, and (4) relevant macroeconomic and microeconomic sentiment. Each model was required to generate one mutually exclusive recommendation—BUY, HOLD, or SELL—and provide its confidence level and analytical rationale. Standardization was implemented to ensure that differences in recommendations primarily reflected differences in model responses rather than variations in information or prompt formulation. The experiment was conducted independently for each stock and trading day. The subsequent trading-day price movement was then used as the benchmark for evaluating the recommendation.

2.5 Operationalization and Measurement of Variables

The principal variables were recommendation accuracy, confidence level, hit ratio, investment return, and investment risk.

2.5.1 Recommendation Accuracy

Recommendation accuracy measures the extent to which an LLM recommendation corresponds to the subsequent direction of stock-price movement. Recommendations were classified into BUY, HOLD, and SELL, while actual market outcomes were categorized as upward, neutral, or downward movements. A BUY recommendation was classified as correct when the subsequent stock price increased, whereas a SELL recommendation was classified as correct when the price decreased. HOLD was considered correct when the subsequent price movement remained within the predetermined neutral range.

2.5.2 Confidence Level

Confidence level represents the degree of certainty stated by each LLM regarding its recommendation. It was treated as a descriptive supplementary measure, rather than an objective probability of prediction accuracy, because the reported percentage was generated by the model and was not statistically calibrated.

2.5.3 Hit Ratio

The hit ratio represents the proportion of profitable BUY recommendations relative to the total BUY recommendations. A higher hit ratio indicates that a larger proportion of BUY signals generated positive subsequent returns.

2.5.4 Investment Return

Investment performance was evaluated using the next-day return generated by BUY recommendations. Simple Cumulative Return (SCR) was used to aggregate individual transaction returns, while Compounded Cumulative Return (CCR) was employed to capture the cumulative effect of sequential reinvestment. HOLD recommendations were treated as non-transactional signals, while SELL recommendations were interpreted as exit signals because the research design did not assume short selling. Transaction costs, taxes, brokerage fees, and dividends were excluded unless otherwise specified.

2.5.5 Investment Risk

Investment risk was evaluated from both predictive and financial-loss perspectives. The analysis incorporated predictive risk, loss rate, total loss, average loss per losing transaction, and a study-specific composite risk score. Predictive risk reflects the proportion of incorrect recommendations, while loss rate represents the proportion of evaluated transactions that generated negative returns. Total loss captures the aggregate magnitude of negative returns, whereas average loss represents the average loss among losing transactions. For comparative purposes, the study constructed a composite risk score by combining predictive risk and loss rate with equal weighting. The resulting score was interpreted as a study-specific comparative indicator, rather than a universally established financial-risk measure.

Table 3.3. Interpretation of the Composite Risk Score

Total Risk Score	Risk Category
0%–25%	Very Low
>25%–50%	Low
>50%–75%	Moderate
>75%–100%	High

2.6 Data Analysis Technique

Data analysis was conducted using Microsoft Excel and IBM SPSS Statistics. The analysis consisted of descriptive statistics, classification-performance evaluation, comparative statistical testing, return analysis, and risk assessment. First, descriptive analysis was performed to determine the frequency and proportion of BUY, HOLD, and SELL recommendations generated by each

LLM. The distribution of AI recommendations was then compared with subsequent actual market conditions. Second, a confusion matrix was used to assess classification performance by comparing LLM recommendations with actual subsequent market conditions. Overall recommendation accuracy was calculated based on the proportion of correctly classified observations. Where appropriate, class-specific measures such as precision, recall, and F1-score were also considered. Third, because BUY, HOLD, and SELL are categorical outcomes, categorical-data procedures were applied rather than conventional normality-based tests. For comparisons involving repeated binary outcomes across the three LLMs, Cochran's Q test was employed. Where a significant difference was identified, pairwise McNemar tests with appropriate multiplicity adjustment were used to determine which models differed significantly. The statistical significance level was set at 5%. A *p*-value below 0.05 indicated statistically significant evidence of differences among the models.

2.7 Return and Risk Evaluation

Return analysis was conducted to determine whether LLM-generated recommendations produced economically meaningful outcomes. The analysis focused primarily on BUY recommendations because the experimental framework did not assume short selling. Three principal measures were employed: hit ratio, Simple Cumulative Return (SCR), and Compounded Cumulative Return (CCR). These measures enabled the study to distinguish between classification performance and actual economic performance. An LLM may have moderate classification accuracy but still generate higher returns if its correct recommendations are associated with relatively large price increases. Risk analysis complemented the return evaluation by examining predictive risk, loss rate, total loss, average loss, and the composite risk score. This approach allowed the study to evaluate whether higher returns were accompanied by greater downside exposure.

2.8 Robustness and Reproducibility

Several procedures were implemented to enhance experimental consistency and reproducibility. First, the same substantive prompt was applied to all three LLMs. Second, equivalent market information was provided for each stock-day observation. Third, each experiment was conducted in a new session to minimize the influence of previous conversational context. Fourth, all AI-generated recommendations were documented immediately after generation. Finally, the coding of BUY, HOLD, and SELL recommendations was standardized before statistical analysis. The experimental dataset was cross-checked against historical market data to minimize transcription errors and ensure consistency in stock identification, observation dates, and price information.

2.9 Research Procedure

The research procedure consisted of ten sequential stages:

1. Sample identification — identifying the 45 LQ45 stocks that met the predetermined criteria.
2. Market-data collection — compiling daily prices, trading volume, candlestick patterns, and relevant economic information.
3. Prompt standardization — developing a uniform investment-analysis prompt.
4. LLM experimentation — submitting equivalent information to ChatGPT, Gemini, and DeepSeek.
5. Recommendation recording — documenting BUY, HOLD, SELL signals, confidence levels, and analytical rationales.
6. Market-outcome determination — identifying the subsequent upward, neutral, or downward price movement.
7. Classification evaluation — calculating recommendation accuracy and related classification measures.

8. Return evaluation — assessing hit ratio, SCR, and CCR.
9. Risk evaluation — assessing predictive risk, loss rate, total loss, average loss, and composite risk.
10. Comparative analysis — statistically comparing the performance of the three LLMs.

Overall, this methodological framework evaluates LLM-based investment recommendations from three complementary dimensions: predictive capability, economic performance, and investment risk. Accordingly, the study treats LLMs as Decision Support Systems (DSS) that assist investors in processing market information rather than as autonomous tools capable of guaranteeing future stock-price movements.

RESULTS AND DISCUSSION

This section presents the empirical findings regarding the ability of Large Language Models (LLMs) to generate stock investment recommendations for companies included in the LQ45 Index of the Indonesia Stock Exchange (IDX). The study evaluates three AI models—ChatGPT, Gemini, and DeepSeek—which were provided with identical prompts containing stock-price information, trading volume, candlestick patterns, and macroeconomic and microeconomic sentiment throughout the observation period. The BUY, HOLD, and SELL recommendations generated by each AI model were subsequently compared with actual stock-price movements to evaluate recommendation accuracy, investment returns, and investment loss risk. Before hypothesis testing was conducted, descriptive statistical analysis was performed to provide an overview of the characteristics of the research data. This analysis was intended to identify the distribution of the data, the investment recommendations generated by the AI models, and the general characteristics of the observations during the study period.

The analysis was followed by a normality test to examine the distribution of the research data and determine the appropriate statistical testing procedure for hypothesis testing. Subsequently, the hypotheses were examined sequentially, beginning with the analysis of LLM recommendation accuracy relative to actual stock-price movements, followed by investment-return analysis and investment-risk analysis for each AI model. The research sample consisted of 45 stocks included in the LQ45 Index during December 2025. The stocks were observed over 20 trading days, resulting in 900 observations for each AI model. Because three AI models—ChatGPT, Gemini, and DeepSeek—were evaluated, the study generated a total of 2,700 AI investment recommendation observations. The research data consisted of BUY, HOLD, and SELL recommendations generated by the AI models based on standardized prompts developed by the researchers using a combination of technical and market-sentiment information. The technical information comprised daily stock prices, trading volume, and candlestick patterns, whereas the sentiment information included relevant macroeconomic conditions and firm-level microeconomic information. All AI-generated investment recommendations were subsequently compared with actual stock-price movements on the following trading day to assess recommendation accuracy, investment returns, and investment loss risk.

3.1.1 Descriptive Statistical Analysis

Descriptive statistical analysis was conducted to provide an overview of the characteristics of the research data. The descriptive statistics included the minimum, maximum, mean, and standard deviation of the relevant research variables. The analysis was intended to identify the general tendencies of investment recommendations generated by each AI model throughout the observation period. In addition, descriptive statistics were used to identify the degree of variation within the research data before conducting further statistical analysis.

3.1.2 Normality Test

A normality test was initially conducted to determine whether the research data followed a normal distribution. The results were intended to provide preliminary information regarding the distributional characteristics of the data and to assist in selecting an appropriate statistical testing procedure. The normality test was conducted using the Kolmogorov–Smirnov and Shapiro–Wilk tests. The decision criteria were as follows:

1. If the significance value (Sig.) is greater than 0.05, the data are considered normally distributed.
2. If the significance value (Sig.) is less than 0.05, the data are considered not normally distributed.

The results of the normality test are presented in Table 4.1.

Table 4.1. Results of the Normality Test

Variable	Kolmogorov–Smirnov Statistic	df	Sig.	Shapiro–Wilk Statistic	df	Sig.
DeepSeek	0.246	898	0.000	0.783	898	0.000
Gemini	0.305	898	0.000	0.734	898	0.000
ChatGPT	0.299	898	0.000	0.750	898	0.000
Actual	0.270	898	0.000	0.771	898	0.000

Source: SPSS output, 2026.

The results presented in Table 4.1 indicate that all variables have significance values of 0.000, which are below the 0.05 threshold. Thus, the null hypothesis of normality is rejected, indicating that the data do not follow a normal distribution. Based on these results, the subsequent analysis employs non-parametric statistical procedures. However, because the core recommendation variables consist of categorical BUY, HOLD, and SELL signals, the non-parametric analysis should be interpreted in accordance with the measurement scale of the variables rather than solely on the basis of normality.

3.1.3 First Hypothesis Testing (H1): LLM Recommendation Accuracy

The first hypothesis examines whether there is a significant difference between the investment signals generated by Large Language Models and the actual subsequent stock-market conditions of LQ45 constituent stocks. The three AI models evaluated in this study are DeepSeek, Gemini, and ChatGPT. The initial analysis considered the distributional characteristics of the data. Because the normality tests indicated non-normal distributions, a non-parametric paired comparison was subsequently conducted.

3.1.3.1 Wilcoxon Signed-Rank Test

The Wilcoxon Signed-Rank Test is a non-parametric procedure designed to evaluate differences between two paired observations. In this study, the test was employed to compare the coded LLM-generated signals with the corresponding coded actual market conditions.
The decision criteria are:

- If ($p < 0.05$), the null hypothesis is rejected, indicating a statistically significant difference.
- If ($p > 0.05$), the null hypothesis is not rejected, indicating insufficient evidence of a statistically significant difference.

The results are presented in Table 4.2.

Table 4.2. Results of the Wilcoxon Signed-Rank Test

Comparison	Z-value	Asymp. Sig. (2-tailed)	Interpretation
Actual – DeepSeek	-7.320	0.000	Significant

Actual – Gemini	-5.362	0.000	Significant
Actual – ChatGPT	-11.800	0.000	Significant

Source: SPSS output, 2026.

The Wilcoxon Signed-Rank Test results indicate that all three AI models produce statistically significant differences relative to the coded actual market conditions. DeepSeek records a Z-value of −7.320 with a significance value of 0.000, Gemini records a Z-value of −5.362 with a significance value of 0.000, and ChatGPT records a Z-value of −11.800 with a significance value of 0.000. Since all significance values are below the 0.05 threshold, the null hypothesis is rejected for all three model comparisons. The results therefore indicate that the distributions of the AI-generated signals differ significantly from the coded actual market conditions during the observation period. Accordingly, H1 is supported, indicating that statistically significant differences exist between the LLM-generated investment signals and the subsequent actual market conditions for LQ45 stocks. Nevertheless, statistical significance should not be interpreted as evidence that the LLM recommendations are necessarily inaccurate or ineffective. A significant difference indicates that the distributions of the coded recommendations and actual market outcomes are statistically different; it does not, by itself, measure the predictive accuracy or economic profitability of an AI model. Therefore, classification metrics and investment-return analysis are required to provide a more complete assessment of model performance.

3.1.4 Second Hypothesis Testing (H2): Investment Return Generated by LLM Signals

The second hypothesis examines whether investment recommendations generated by Large Language Models can produce positive investment returns for LQ45 stocks during the observation period. The investment-return analysis evaluates the economic consequences of following the recommendations generated by DeepSeek, Gemini, and ChatGPT. The analysis employs two complementary measures: Simple Cumulative Return (SCR) and Compounded Cumulative Return (CCR). Because the study does not assume short-selling positions, the primary return analysis focuses on BUY recommendations. HOLD recommendations are interpreted as decisions not to initiate a new position, while SELL recommendations are interpreted as exit signals rather than short-selling positions. The return generated by each BUY recommendation is calculated by comparing the stock price at the time of the recommendation with the stock price on the subsequent trading day. The individual transaction returns are then aggregated to obtain the SCR and CCR for each AI model. The use of both measures provides a more comprehensive assessment of investment performance. SCR captures the aggregate contribution of individual transaction returns, whereas CCR reflects the growth of capital under a sequential reinvestment assumption. The resulting return performance of DeepSeek, Gemini, and ChatGPT is subsequently compared to determine which model generates the most favorable economic outcome during the December 2025 observation period.

Table 4.3. Results of LLM Recommendations (DeepSeek) and Actual Stock Prices

DEEPSEEK	ACTUAL STOCK PRICE	DESCRIPTION
Prediction Date	Stock	AI Recommendation

Table 4.3. Results of LLM Recommendations (DeepSeek) and Actual Stock Prices

[The data values remain unchanged from the original table.]
Note: *BUY, HOLD, and SELL represent investment recommendations generated by DeepSeek. "Correct" indicates that the recommendation was consistent with the subsequent direction of actual stock price movements;*

"Incorrect" indicates inconsistency; and "Neutral" indicates a HOLD recommendation that does not imply a directional prediction.

Based on the results of the LQ45 stock data analysis, particularly banking stocks, DeepSeek provided investment recommendations in the form of buy, hold, and sell based on the information provided in the research prompt. This information was used by DeepSeek to generate investment decisions for each observation date. The resulting recommendations were subsequently compared with actual stock price movements during the period following the issuance of each recommendation. The data analysis shows that DeepSeek generated 24 recommendations categorized as correct, 30 as incorrect, and 46 as neutral, out of 100 research observations. The correct category indicates that the recommendation generated by DeepSeek was consistent with the direction of the subsequent actual stock price movement. The incorrect category indicates that the DeepSeek recommendation was inconsistent with the actual price movement, whereas the neutral category represents situations in which a HOLD recommendation was not interpreted as indicating either an upward or downward price movement.

These results indicate that DeepSeek was not consistently able to follow the direction of actual stock price movements. The number of incorrect recommendations remained higher than the number of correct recommendations. Nevertheless, the relatively high number of neutral recommendations indicates that DeepSeek frequently generated HOLD recommendations when market conditions did not provide sufficiently strong signals to determine the direction of future price movements. The findings suggest that DeepSeek was capable of generating investment recommendations based on the market information provided; however, these recommendations did not always correspond to the subsequent direction of actual market movements. Such differences may occur because stock prices are affected by changes in market conditions that emerge after the recommendation has been generated, whereas DeepSeek's recommendations are based on the information available at the time the decision is made.

The similarity between LLM returns and actual stock price returns in several observations does not indicate that the Large Language Model (LLM) generated or knew the actual stock prices. Rather, this similarity occurs because the LLM return in this study is calculated based on actual stock price changes following the issuance of an investment recommendation. Thus, the LLM functions as a provider of investment signals or decisions, while the return value is derived from the realized changes in market prices. For example, if DeepSeek provides a BUY recommendation on a particular date when the stock price is IDR 8,350 and the stock price subsequently increases to IDR 8,375, the actual return is calculated based on this price change. Because the LLM return in this study is calculated using the realized price change following a BUY recommendation, the LLM return and actual stock price return may have identical values. Therefore, such similarity does not imply that the price generated by the LLM is identical to the market price.

The LLM does not determine the actual stock price. It only provides a transaction recommendation, whereas the actual stock price is the price formed in the Indonesia Stock Exchange. Accordingly, in this study, the LLM is evaluated based on the accuracy of its recommendations relative to realized stock price movements, rather than on its ability to generate stock prices identical to market prices.

Table 4.4. Results of LLM Recommendations (Gemini) and Actual Stock Prices

GEMINI	ACTUAL STOCK PRICE	DESCRIPTION
Prediction Date	Stock	AI Recommendation

Table 4.4. Results of LLM Recommendations (Gemini) and Actual Stock Prices

Note: *BUY, HOLD, and SELL represent investment recommendations generated by Gemini. "Correct" indicates that the recommendation was consistent with the subsequent direction of actual stock price movements; "Incorrect" indicates inconsistency; and "Neutral" indicates a HOLD recommendation that does not imply a directional prediction.*

Based on the results of the research data analysis, Gemini provided stock investment recommendations in the form of buy, hold, and sell based on market information supplied through the research prompt. These recommendations were subsequently compared with actual stock price movements to assess the consistency between the investment decisions generated by Gemini and actual market conditions. Of the 100 research observations, Gemini generated 29 recommendations categorized as correct, 32 as incorrect, and 39 as neutral. These findings indicate that Gemini generated the highest number of correct recommendations compared with the number of correct recommendations generated by DeepSeek and ChatGPT in the research dataset. Nevertheless, the number of incorrect recommendations generated by Gemini remained higher than the number of correct recommendations. This condition indicates that Gemini was not consistently able to provide recommendations that corresponded to the direction of actual stock price movements. Meanwhile, the 39 neutral recommendations indicate that Gemini relatively frequently generated HOLD decisions when market conditions did not provide sufficiently strong signals for either BUY or SELL transactions.

These findings suggest that Gemini has the capability to identify market conditions based on the information provided; however, limitations remain in consistently determining the direction of subsequent stock price movements. The differences between Gemini's recommendations and actual market conditions constitute an important component of evaluating the capability of LLMs as Decision Support Systems (DSS) in this study. The returns used to evaluate Gemini do not represent stock prices predicted by Gemini. Instead, the return values are derived from actual changes in stock prices following the issuance of the investment recommendations. Therefore, when the LLM return and actual stock price return have identical values in a particular observation, this indicates that both measures are calculated using the same realized stock price changes as their underlying basis. Within the context of this study, Gemini functions as a system that generates BUY, HOLD, or SELL decisions, whereas actual stock prices are used to evaluate the consequences of those decisions. Consequently, identical return values should not be interpreted as evidence that Gemini accurately predicted stock prices. Rather, such similarity simply indicates that the returns were realized using the same actual stock price data.

The subsequent performance of Gemini was evaluated based on whether the recommendations were consistent with the direction of actual stock price movements and whether the resulting decisions generated positive or negative returns. Under this approach, Gemini's capability is assessed as a decision-support system rather than as a system capable of generating stock prices that substitute for actual market prices.

Table 4.4. Results of LLM (Gemini) Recommendations and Actual Stock Prices

Prediction Date	Stock	AI Recommendation	Confidence Level (%)	Date	Open	High	Low	Close	Volume	LLM Return	Actual Price Return	CHAT GPT	ChatGPT Code
01/12/2025	BBCA	HOLD	75	Dec 1, 2025	8,450	8,450	8,350	8,400	76,359,200			Neutral	2

02/12/2025	BBCA	HOLD	65	Dec 2, 2025	8,350	8,425	8,350	8,375	97,044,000			Neutral	2
03/12/2025	BBCA	BUY	60	Dec 3, 2025	8,350	8,400	8,300	8,300	103,550,900	0.01	0.01	Incorrect	0
04/12/2025	BBCA	HOLD	55	Dec 4, 2025	8,325	8,350	8,225	8,225	71,410,400			Neutral	2
05/12/2025	BBCA	HOLD	65	Dec 5, 2025	8,200	8,300	8,200	8,300	71,162,300			Neutral	2
08/12/2025	BBCA	HOLD	65	Dec 8, 2025	8,350	8,350	8,250	8,300	61,745,200			Neutral	2
09/12/2025	BBCA	HOLD	70	Dec 9, 2025	8,300	8,300	8,100	8,100	113,057,200			Neutral	2
10/12/2025	BBCA	HOLD	55	Dec 10, 2025	8,100	8,175	8,025	8,075	143,418,300			Neutral	2
11/12/2025	BBCA	HOLD	75	Dec 11, 2025	8,175	8,200	7,900	8,000	140,766,000			Neutral	2
12/12/2025	BBCA	HOLD	55	Dec 12, 2025	7,900	8,000	7,850	8,000	100,006,800			Neutral	2
15/12/2025	BBCA	BUY	72	Dec 15, 2025	7,925	8,300	7,925	8,300	143,105,700	-0.04	-0.04	Neutral	2
16/12/2025	BBCA	BUY	82	Dec 16, 2025	8,375	8,375	8,025	8,075	122,332,000	0.03	0.03	Incorrect	0
17/12/2025	BBCA	HOLD/BUY	75	Dec 17, 2025	8,000	8,125	7,975	8,025	89,802,700	0.02	0.02	Neutral	2
18/12/2025	BBCA	HOLD	70	Dec 18, 2025	8,025	8,250	8,025	8,175	88,497,200			Neutral	2
19/12/2025	BBCA	BUY	78	Dec 19, 2025	8,200	8,225	8,050	8,050	131,069,800	0.02	0.02	Incorrect	0
22/12/2025	BBCA	HOLD	65	Dec 22,	8,050	8,175	8,025	8,175	72,424,200			Neutral	2

				2025									
23/12/2025	BBCA	HOLD/BUY	78	Dec 23, 2025	8,100	8,125	7,975	8,025	85,044,100	0.02	0.02	Neutral	2
24/12/2025	BBCA	HOLD	70	Dec 24, 2025	8,025	8,075	8,000	8,025	71,260,500			Neutral	2
29/12/2025	BBCA	HOLD	70	Dec 29, 2025	8,025	8,025	7,950	8,025	78,399,400			Neutral	2
30/12/2025	BBCA	HOLD	60	Dec 30, 2025	7,950	8,175	7,950	8,075	101,995,600			Neutral	2

Note: The remaining observations for BBRI, BBNI, BBTN, and BMRI follow the same translation convention: Benar → Correct, Salah → Incorrect, Netral → Neutral, Harga Aktual → Actual Price, Rata-Rata Return → Average Return, Tingkat Keyakinan → Confidence Level, Prediksi untuk Tanggal → Prediction Date, and Kode ChatGPT → ChatGPT Code.

Results and Discussion

Based on the results of the data analysis, ChatGPT provided stock investment recommendations in the form of BUY, HOLD, and SELL based on the information available in the research prompt. These recommendations were subsequently compared with actual stock price movements to determine the degree of consistency between the investment decisions generated by the model and actual market conditions. From the 100 research observations, ChatGPT generated 15 correct recommendations, 27 incorrect recommendations, and 58 neutral recommendations. These results indicate that ChatGPT produced a higher number of neutral recommendations than DeepSeek and Gemini. The high proportion of neutral recommendations indicates ChatGPT's tendency to issue HOLD decisions under various market conditions. This strategy may be viewed as a relatively conservative approach because the model does not actively issue BUY or SELL recommendations when the direction of price movement is not sufficiently strong based on the available information.

Nevertheless, the number of correct ChatGPT recommendations was lower than the number of incorrect recommendations. This finding indicates that the decisions generated by ChatGPT did not consistently correspond with subsequent stock price realizations. Therefore, the findings suggest that ChatGPT can function as a decision-support tool in the investment process, but its recommendations remain subject to prediction errors when evaluated against actual market conditions. Similar to the findings for DeepSeek and Gemini, the LLM Return reported for ChatGPT in this study does not represent a stock price predicted by ChatGPT. Instead, the return is calculated based on the actual change in stock prices after the recommendation was issued. Therefore, when the LLM Return and Actual Price Return have identical values, this similarity results from the use of the same actual stock prices in the return calculation.

ChatGPT only serves as a provider of investment decision signals, whereas the financial outcome of those decisions is determined by the actual stock price movements observed in the market. When a BUY recommendation is issued and the stock price subsequently increases, the decision generates a positive return. Conversely, when the stock price declines, the decision generates a negative return. Thus, the evaluation of ChatGPT in this study is not based on whether

ChatGPT can generate a price identical to the market price, but rather on whether its investment decisions correspond to actual market movements and generate the expected return consequences. Based on the data analysis, DeepSeek generated 24 correct, 30 incorrect, and 46 neutral recommendations, Gemini generated 29 correct, 32 incorrect, and 39 neutral recommendations, while ChatGPT generated 15 correct, 27 incorrect, and 58 neutral recommendations. These findings indicate that the three LLMs exhibit different recommendation characteristics and have not yet demonstrated a fully consistent ability to track actual stock price movements. Differences in the numbers of correct, incorrect, and neutral recommendations indicate that the accuracy of investment decisions varied across the models throughout the research period.

Furthermore, to examine the financial consequences of these recommendations, an analysis of the cumulative returns generated by the LLM recommendations and actual stock prices was conducted. This analysis was intended to examine changes in investment value based on the recommendations generated by each model and the actual stock price realizations in the market. In this study, LLMs function as providers of BUY, HOLD, or SELL signals, whereas actual stock prices serve as the basis for calculating returns. Therefore, when LLM Return and Actual Price Return have identical values, this occurs because both measures are calculated using actual price changes, rather than because the LLM generated a stock price identical to the market price.

Table 4.6. Cumulative Returns of LLM Models and Actual Stock Prices

AI Model	Total Observations	Positive Returns	Negative Returns	SCR LLM	SCR Actual Stock	CCR LLM	CCR Actual Stock
DeepSeek	900	610	290	4.92%	4.92%	24.01%	24.01%
Gemini	900	598	302	4.86%	4.86%	23.58%	23.58%
ChatGPT	900	470	430	1.20%	1.20%	1.90%	1.90%

Source: Data processed by the researchers (2026).

Table 4.6 shows that all Large Language Models (LLMs) generated positive investment returns. The study examined 45 LQ45 stocks over 20 trading days, resulting in 900 observations for each AI model. DeepSeek generated the highest Compounded Cumulative Return (CCR) of 24.01%, followed by Gemini at 23.58%, while ChatGPT generated a CCR of 1.90%. These findings indicate that investment recommendations generated by AI models have the potential to generate positive stock investment returns during the research period. Furthermore, the number of positive returns generated by DeepSeek and Gemini was higher than their negative returns. This finding suggests that the BUY, HOLD, and SELL recommendations generated by these two AI models were more frequently aligned with actual stock price movements in the capital market.

Table 4.7. Interpretation of Cumulative Returns

AI Model	SCR	CCR	Interpretation
DeepSeek	4.92%	24.01%	Very Good
Gemini	4.86%	23.58%	Good
ChatGPT	1.20%	1.90%	Fair

Source: Data processed by the researchers (2026).

Table 4.7 presents an interpretation of the effectiveness of each AI model in generating stock investment returns. DeepSeek demonstrated the strongest performance because it generated the highest CCR value, indicating a relatively stronger ability to identify stock price movements during the research period. Gemini also demonstrated strong performance, with a return value close to that of DeepSeek. Meanwhile, ChatGPT continued to generate a positive return, although its value was considerably lower than those of the other AI models. This finding indicates that all AI models examined in this study retained some capacity to generate positive investment returns. Overall, the Large Language Models (LLMs) examined in this study generated positive investment returns based on the Simple Cumulative Return (SCR) and Compounded Cumulative Return

(CCR) calculations. DeepSeek generated a CCR of 24.01%, Gemini generated 23.58%, and ChatGPT generated 1.90%.

The positive returns were obtained by comparing the BUY, HOLD, and SELL recommendations generated by the AI models with actual stock price movements for LQ45 stocks during the research period. These findings indicate that investment recommendations based on Large Language Models (LLMs) have the potential to generate positive stock investment returns. Therefore, Hypothesis 2 (H2) is accepted, indicating that H2 predicts a positive investment return.

Investment Loss Risk

The third hypothesis (H3) aims to measure the level of investment loss risk associated with stock recommendations generated by Large Language Models (LLMs), namely OpenAI ChatGPT, Google Gemini, and DeepSeek. Investment loss risk in this study was measured using predictive risk, financial risk, average loss per losing trade, loss rate, and total risk. These indicators were employed to assess the magnitude of potential losses that investors may experience when following investment recommendations generated by each AI model. The measurement of investment loss risk was conducted to determine the extent to which investment recommendations generated by LLMs can minimize investors' potential losses in stock transactions. Investment risk was analyzed using several risk indicators reflecting prediction errors and actual financial losses generated by each AI model. The use of predictive risk, loss rate, and financial risk indicators is consistent with the downside-risk perspective in investment analysis, which focuses on measuring investors' potential losses.

Stock returns were calculated using the daily change in stock prices following the issuance of AI recommendations. Stock returns were calculated using the following formula: When the return value is below zero (< 0), the return is categorized as a negative return or investment loss. Negative returns were subsequently used to calculate the risk level associated with each AI model. In addition, this study evaluated the accuracy of AI recommendations relative to the direction of actual stock price movements. The evaluation employed three categories: correct, incorrect, and neutral. A recommendation was categorized as correct when the AI recommendation was consistent with the direction of actual stock price movements, whereas it was categorized as incorrect when the AI recommendation was inconsistent with actual price movements. HOLD recommendations were categorized as neutral. Based on the results of the data analysis for LQ45 stocks during December 2025, the investment risk associated with each AI model was obtained as follows.

Table 4.8. Investor Loss Risk

Risk Variable	DeepSeek	Actual Price	Gemini	Actual Price	ChatGPT	Actual Price
Correct	257	—	230	—	185	—
Incorrect	307	—	341	—	246	—
Neutral	336	—	329	—	469	—
Risk Accuracy	46%	—	40%	—	43%	—
Total Negative Return	238.00	238.00	170.00	170.00	115.00	115.00
Predictive Risk	54.43%	54.43%	61.00%	61.00%	37.61%	37.61%
Financial Risk	0.0492	0.0492	0.0485	0.0485	0.0119	0.0119
Average Loss per Losing Trade	0.7752	0.7752	0.4985	0.4985	0.4674	0.4674
Loss Rate	34.11%	34.11%	37.88%	37.88%	27.33%	27.33%
Total Risk	22.34%	22.34%	24.85%	24.85%	16.35%	16.35%

Based on these findings, Gemini produced the highest predictive risk, at 61.00%, with a recommendation accuracy of 40%, which falls within the low-accuracy category, and 170.00 negative-return observations. This was followed by DeepSeek, with a predictive risk of 54.43%

and recommendation accuracy of 46%, also categorized as low, with 238.00 negative-return observations. Meanwhile, ChatGPT produced the lowest predictive risk, at 37.61%, with recommendation accuracy of 43%, also categorized as low, and 115.00 negative-return observations. These findings indicate that ChatGPT demonstrated a relatively better ability to predict the direction of stock prices than Gemini and DeepSeek. For the financial risk indicator, DeepSeek produced a value of 0.0492, Gemini 0.0485, and ChatGPT 0.0119. These values indicate that ChatGPT generated the lowest level of actual financial loss compared with the other AI models.

Furthermore, the average loss per losing trade shows that DeepSeek generated the largest average loss per losing transaction, at 0.7752, followed by Gemini at 0.4985 and ChatGPT at 0.4674. The findings also show that Gemini had the highest loss rate, at 37.88%, followed by DeepSeek at 34.11%, while ChatGPT generated the lowest loss rate, at 27.33%. A lower loss rate indicates a lower level of investment loss risk associated with the AI-generated recommendations. Based on overall total risk, Gemini generated the highest risk level, at 24.85%, followed by DeepSeek at 22.34%, while ChatGPT generated the lowest total risk, at 16.35%. These findings indicate that ChatGPT was the AI model associated with the lowest level of investment risk during the research period. Overall, ChatGPT demonstrated the strongest risk performance compared with Gemini and DeepSeek because it generated the lowest values for predictive risk, financial risk, loss rate, and total risk. Conversely, Gemini exhibited the highest overall investment risk during the research period.

The findings also demonstrate that all AI models continued to generate negative returns in several observations. This indicates that AI cannot completely eliminate investment risk because stock markets remain affected by market volatility, investor sentiment, macroeconomic and microeconomic conditions, and continuously changing market information. From an academic perspective, these findings support the Decision Support System (DSS) theory, which explains that artificial intelligence can assist investors in making investment decisions by processing market data and information rapidly and systematically. The findings also support bounded rationality theory, which explains that investors face limitations in processing complex market information; therefore, the use of AI can help mitigate such limitations. The findings are also consistent with Dowling and Lucey (2023) and Rao and Patel (2023), who explain that AI models may still be subject to prediction bias and overconfidence under certain market conditions. In addition, the findings support Pelster and Val (2024), who found that LLM-based recommendations can assist investors in identifying market direction but remain subject to prediction errors due to highly volatile changes in market conditions.

The decision criteria for the third hypothesis (H3) were determined based on the level of investment loss risk generated by each AI model. H3 is accepted when investment risk falls within the low-to-moderate category and the loss rate is below 50%. Conversely, H3 is rejected when investment risk falls within the high-risk category and the loss rate exceeds 50%. A loss rate below 50% indicates that loss-making transactions do not dominate the overall investment transactions; therefore, investment risk remains within the low-to-moderate category.

Based on the research findings:

- DeepSeek had a loss rate of 34.11%;
- Gemini had a loss rate of 37.88%; and
- ChatGPT had a loss rate of 27.33%.

All AI models had loss rates below 50%, while their total risk levels remained within the low-to-moderate category. Furthermore, loss-making transactions did not dominate the overall research observations. This indicates that investment recommendations based on Large Language Models (LLMs) may assist investors in reducing the risk associated with stock investment decisions. Overall, the findings demonstrate that Large Language Models (LLMs) have the capacity

to assist investors in reducing the risk associated with stock investment decisions, although they cannot completely eliminate market risk. Therefore, based on the analysis of predictive risk, financial risk, average loss per losing trade, loss rate, and total risk, as well as the predetermined decision criteria for the hypothesis, the research hypothesis can be concluded as follows: H3 is accepted because the signals generated by Large Language Models (LLMs) resulted in a relatively low level of investment loss risk in stock investment.

DISCUSSION

This study aims to analyze the ability of Large Language Models (LLMs), namely ChatGPT, Gemini, and DeepSeek, to provide stock investment recommendations for LQ45 stocks listed on the Indonesia Stock Exchange (IDX). The three models were provided with the same prompt incorporating stock price information, trading volume, candlestick patterns, and macroeconomic and microeconomic sentiment. The evaluation was conducted across three main aspects corresponding to the research hypotheses: the difference between LLM signals and actual stock prices, the ability to generate positive returns, and the level of investment loss risk. The findings were subsequently interpreted by linking the empirical results to the theories of Bounded Rationality, Prospect Theory, the Efficient Market Hypothesis (EMH), and the Decision Support System (DSS).

a. Discussion of H1: Differences Between LLM Signals and Actual Stock Prices

The results of the Wilcoxon Signed-Rank Test indicate a significant difference between the signals generated by DeepSeek, Gemini, and ChatGPT and the actual stock price conditions. The significance value for all models was 0.000, which is below 0.05; therefore, H1 is accepted. This finding indicates that LLM recommendations are not yet capable of perfectly representing actual stock price movements. In other words, the ability of LLMs to interpret market information does not imply that every Buy, Hold, or Sell recommendation will always correspond to the direction of stock prices in the subsequent period. The findings demonstrate a discrepancy between the recommendation signals generated by LLMs and actual stock price movements during the research period. This discrepancy is reflected in the accuracy levels of the respective AI models, which were unable to predict stock price direction perfectly. Although AI possesses the capability to process large amounts of market data and information, AI models remain subject to limitations in predicting complex and volatile stock market dynamics. This discrepancy can be explained through the Bounded Rationality Theory. Investors face limitations in processing large volumes of complex and rapidly changing information. LLMs can help reduce these limitations because they are capable of systematically processing price data, trading volume, candlestick patterns, and market sentiment. Nevertheless, LLMs remain constrained because their recommendations depend on the information available and the patterns that can be identified by the models. Therefore, the H1 findings demonstrate that AI can support the analytical process but cannot eliminate market uncertainty or guarantee accurate price direction predictions.

The H1 findings can also be associated with the Efficient Market Hypothesis (EMH). Available market information can be rapidly incorporated into stock prices, while stock prices may also change as a result of new information emerging after a recommendation has been generated. This condition explains why LLM signals do not always correspond to actual price realizations. Accordingly, the findings should not be interpreted as indicating that LLMs fail entirely; rather, they demonstrate that AI-based information processing remains subject to market uncertainty. Within the Decision Support System (DSS) framework, LLMs are more appropriately positioned as supporting tools that provide information and signals to assist investors in decision-making rather than as instruments capable of providing certainty regarding future price movements. These findings are consistent with J. Chen et al. (2023), who found that AI-based models are capable of generating significant predictions of daily return direction, although their accuracy fluctuates depending on market conditions. In addition, Martinez et al. (2026) reported that GPT-4

demonstrated higher accuracy than other AI models, although its accuracy remained at a moderate level and therefore could not fully replace investor analysis. These findings are further supported by Achmadi, Haanurat, and Rustam (2020), who found that stock trading volume influences stock prices. Their study, which examined stocks included in the Jakarta Islamic Index, demonstrated that trading volume is one of the factors associated with stock prices. This finding is relevant to the present study because trading volume constitutes one of the information components incorporated into the LLM prompts. Thus, stock price movements are not determined by a single information variable but may be influenced by multiple interacting factors. This helps explain why LLM signals may differ from actual stock market conditions.

Differences in recommendations across AI models also indicate that each model possesses distinct data-processing characteristics. ChatGPT tends to provide more stable recommendations, supported by its natural language processing capabilities, whereas Gemini and DeepSeek demonstrate greater variation and may be more sensitive to changes in market information. This condition suggests that differences in AI model architecture may influence the ability of each model to interpret technical data and market sentiment. Based on the statistical testing results, the first hypothesis is accepted. Therefore, it can be concluded that there is a significant difference between the signals generated by LLMs and actual stock prices in stock investment.

b. Discussion of H2: Stock Investment Returns

The findings indicate that all LLM models generated positive investment returns during the research period. DeepSeek generated a Compounded Cumulative Return (CCR) of 24.01%, followed by Gemini at 23.58%, while ChatGPT generated a CCR of 1.90%. Therefore, H2 is accepted because the LLM signals were able to generate positive cumulative returns during the observation period. DeepSeek produced the highest return, followed by Gemini, while ChatGPT also generated a positive return, although at a relatively lower level. These findings demonstrate that the difference between LLM signals and actual stock prices identified in H1 does not automatically result in negative investment performance. This finding is important because prediction accuracy and profitability are not necessarily identical. A model may generate incorrect recommendations in some observations while still producing a positive cumulative return if gains from successful recommendations are sufficient to offset losses from unsuccessful recommendations. Therefore, the performance of LLMs in this study should be assessed not only based on recommendation accuracy but also based on their economic consequences for investment returns.

The H2 findings can be explained through the Decision Support System (DSS) perspective. LLMs are capable of integrating various types of information into Buy, Hold, and Sell signals that can be considered by investors. The positive returns generated during the research period indicate that the information processed by LLMs possesses economic value in supporting investment decisions. However, because these results were obtained over a limited research period, positive returns cannot be interpreted as evidence that the models will consistently generate profits under different market conditions. The findings demonstrate that investment recommendations generated by LLMs were capable of producing positive investment returns during the research period. Positive returns indicate that AI possesses a certain ability to interpret stock price movements and potentially provide investment opportunities when Buy, Hold, and Sell recommendations are followed systematically.

The positive returns observed in this study are consistent with Pelster and Val (2024), who found that LLM-based recommendations are associated with short-term abnormal returns. Furthermore, Ko, Kim, et al. (2024) reported that ChatGPT-based investment recommendations could improve portfolio performance and compete with market benchmarks. These findings are also relevant to Khaerunnisa, Rustam, and Rustan (2024) in *The Influence of Financial Literacy, Financial Efficacy and Demographic Factors on Investment Decisions in the Capital Market*, which demonstrated that financial literacy and financial efficacy have positive and significant effects on

investment decisions in the capital market. This evidence reinforces the argument that information quality and the ability to understand financial information constitute important components of investment decision-making. In the present study, LLMs can therefore be regarded as supporting tools that assist in systematically processing market information, although their recommendations must still be understood and evaluated by investors before being used as the basis for investment decisions.

The ability of AI to generate positive returns is influenced by the combination of technical analysis and market sentiment incorporated into the research prompt. Information concerning candlestick patterns, trading volume, macroeconomic conditions, and firm-level sentiment enables AI models to develop a more comprehensive understanding of market momentum. When technical conditions and sentiment are supportive, AI models tend to generate more accurate Buy recommendations, thereby creating opportunities for positive returns. Nevertheless, returns were not consistently positive for every stock or every trading day. Under volatile market conditions, several AI recommendations generated negative returns because of rapid changes in market sentiment. This finding demonstrates that AI remains limited in its ability to predict all dimensions of stock market dynamics accurately. From the perspective of the Efficient Market Hypothesis (EMH), the findings indicate that AI is still capable of identifying potential profit opportunities from available market information. This condition suggests that the Indonesian capital market may not be fully efficient in the semi-strong form, thereby leaving opportunities for AI models to identify market signals that are not yet fully reflected in stock prices. Based on the findings, the second hypothesis is accepted. Therefore, it can be concluded that LLM signals are capable of generating positive investment returns in stock investment.

c. Discussion of H3: Stock Investment Loss Risk

The findings indicate that all three LLM models exhibited investment loss risks within the low-to-moderate category based on the criteria established in this study. ChatGPT generated the lowest total risk at 16.35%, followed by DeepSeek at 22.34% and Gemini at 24.85%. In terms of predictive risk, ChatGPT also recorded the lowest value at 37.61%, compared with DeepSeek at 54.43% and Gemini at 61.00%. ChatGPT recorded a loss rate of 27.33%, DeepSeek 34.11%, and Gemini 37.88%. Thus, all models recorded loss rates below the 50% threshold established for H3. These findings demonstrate a trade-off between return and risk. DeepSeek generated the highest return but was not the model with the lowest risk. Conversely, ChatGPT generated a lower return than DeepSeek and Gemini but exhibited the lowest predictive risk, financial risk, average loss per losing trade, loss rate, and total risk. This finding demonstrates that the AI model generating the highest return is not necessarily the best model when evaluated from a risk perspective. The findings are further supported by Bounded Rationality Theory and the DSS framework. Investors face limitations in simultaneously processing diverse market information, whereas LLMs can assist in systematically processing price, volume, candlestick, and economic sentiment information. Nevertheless, risk remains because market conditions change dynamically and not all information can be predicted by the model. Therefore, H3 should be interpreted as indicating that LLMs can assist investors in reducing decision-making limitations and managing risk-related information, but cannot eliminate investment risk.

For investors, selecting AI-generated recommendations should consider the balance between potential returns and potential losses. Therefore, H3 is accepted, indicating that investment recommendations generated by LLMs still involve investment loss risk, although the level of risk remains within a range that can be tolerated by investors. In this study, investment risk was measured using the loss rate, percentage of negative returns, and maximum drawdown to assess the potential losses associated with AI-generated recommendations. The findings demonstrate that although AI can generate positive returns, several recommendations still resulted in losses because actual market movements diverged from the predictions generated by the AI models. This condition demonstrates that AI cannot completely eliminate investment risk because

stock markets are characterized by substantial uncertainty and are influenced by external factors such as global sentiment, economic policies, and investor behavior. These findings are consistent with Dowling and Lucey (2023), who explained that LLM-based models continue to face prediction errors, particularly under highly volatile market conditions. Furthermore, S. Rao and Patel (2023) found that AI models may exhibit overconfidence when generating recommendations, thereby increasing the risk of incorrect investment decisions.

Differences in risk levels across AI models indicate that each model has a different degree of sensitivity to changes in market conditions. AI models with higher predictive accuracy tend to exhibit lower loss rates and maximum drawdowns than other models. Thus, investment risk is strongly influenced by the ability of an AI model to interpret market momentum and investor sentiment. From the perspective of Prospect Theory, the findings demonstrate that investors still need to consider the psychological risks associated with using AI-generated recommendations. Although AI can assist investors in decision-making, investors may still experience loss aversion when facing investment losses resulting from recommendations that do not correspond to actual market conditions. Based on the findings, the third hypothesis is accepted. Therefore, it can be concluded that LLM signals generate investment loss risks that remain within a tolerable range for stock investment. Overall, the three hypotheses demonstrate that the capabilities of LLMs as Decision Support Systems are multidimensional. H1 demonstrates a significant difference between LLM signals and actual market conditions; H2 demonstrates that despite prediction errors, LLM recommendations can still generate positive cumulative returns; while H3 demonstrates that investment loss risk remains within the low-to-moderate category according to the research criteria. Therefore, LLMs are more appropriately positioned as information-based investment decision-support tools that assist investors in processing market information rather than as substitutes for investors or instruments capable of providing certainty regarding future stock price movements.

D. Research Limitations

This study has several limitations that should be considered when interpreting the findings. First, the research period was relatively short, covering only 20 trading days in December 2025. Consequently, the findings do not fully capture the consistency of AI model performance across different long-term market conditions. Second, this study focused exclusively on stocks included in the LQ45 Index. Therefore, the findings cannot yet be generalized to all stocks listed on the Indonesia Stock Exchange, particularly stocks with lower liquidity or different volatility characteristics. Third, the investment recommendations generated by AI models are strongly influenced by the quality and structure of the prompts provided. Although this study employed standardized prompts, potential model interpretation bias toward the prompts may still occur and influence the resulting Buy, Hold, and Sell recommendations. Fourth, this study employed Large Language Models (LLMs), which are dynamic systems that may undergo periodic updates.

Changes in model versions, system parameters, and AI capabilities over time may produce different recommendation outputs even when identical prompts are used. Fifth, this study primarily employed a combination of technical data and macroeconomic and microeconomic sentiment without incorporating firm-level fundamental variables in depth. Consequently, the AI-generated recommendations may not fully reflect the overall fundamental conditions of the respective issuers. Sixth, this study focused on evaluating the accuracy, returns, and risks associated with AI-generated recommendations but did not examine investor psychology, decision-making behavior, or other external factors such as geopolitical conditions and changes in global economic policies, which may significantly affect stock markets. Therefore, future research is encouraged to employ a longer observation period, incorporate firm-level fundamental variables, expand the research sample beyond LQ45 stocks, and compare a broader range of AI models. Such extensions would provide more comprehensive findings and improve the generalizability of the research results.

CONCLUSION

A. Conclusions

This study examined the ability of Large Language Models (LLMs), namely ChatGPT, Gemini, and DeepSeek, to generate stock investment recommendations for stocks included in the LQ45 Index of the Indonesia Stock Exchange (IDX). The evaluation was conducted by comparing the Buy, Hold, and Sell signals generated by each model with actual stock price movements, investment returns, and investment loss risks during the research period. Based on the statistical tests and empirical analysis, several conclusions can be drawn. First, there is a significant difference between the investment signals generated by LLMs and actual stock price movements. The findings indicate that the Buy, Hold, and Sell recommendations generated by ChatGPT, Gemini, and DeepSeek do not always correspond to the subsequent direction of actual stock prices. This discrepancy suggests that the predictive capabilities of LLMs remain subject to market volatility, changes in investor sentiment, new information emerging after recommendations are generated, and the inherent limitations of AI models in capturing all factors affecting stock price movements.

Nevertheless, LLMs demonstrate potential as decision support systems for investment analysis rather than as deterministic stock price prediction instruments. Second, LLM-generated investment signals were capable of producing positive cumulative investment returns during the research period. DeepSeek generated the highest Compounded Cumulative Return (CCR) of 24.01%, followed by Gemini at 23.58%, while ChatGPT generated a CCR of 1.90%. These findings indicate that although some recommendations were inaccurate, successful recommendations were sufficient to generate positive cumulative returns. Therefore, prediction accuracy and investment profitability should not necessarily be regarded as identical dimensions of AI performance. Third, LLM-generated recommendations produced relatively low-to-moderate investment loss risk based on the criteria established in this study. ChatGPT recorded the lowest total risk at 16.35%, followed by DeepSeek at 22.34% and Gemini at 24.85%. Similarly, the loss rates were 27.33% for ChatGPT, 34.11% for DeepSeek, and 37.88% for Gemini, all below the 50% threshold established in the research criteria. These findings indicate that LLMs cannot eliminate investment risk but may function as supporting instruments for identifying investment opportunities and managing decision-making risks.

Fourth, the three AI models demonstrated different performance characteristics in generating stock investment recommendations. DeepSeek produced the highest investment return, Gemini generated a return close to that of DeepSeek, whereas ChatGPT demonstrated the lowest level of investment risk. These differences suggest that each model has distinct capabilities in processing technical information, trading volume, candlestick patterns, and economic sentiment. Therefore, no single AI model can be considered universally superior across all dimensions of investment performance. Overall, this study demonstrates that LLMs possess analytical and economic value as decision support systems for stock investment. However, they cannot replace investors or provide certainty regarding future stock price movements. LLMs can facilitate information processing and generate investment signals, while final investment decisions should remain subject to fundamental analysis, technical analysis, market conditions, and appropriate risk management.

B. Theoretical Implications

This study provides several theoretical implications for the growing literature on artificial intelligence and investment decision-making. First, the findings support the perspective of Bounded Rationality Theory, which suggests that investors face limitations in processing large volumes of complex and rapidly changing information. LLMs may reduce these limitations by integrating diverse market information into systematic investment signals. However, the findings also demonstrate that AI cannot eliminate information constraints and market uncertainty. Second, this study contributes to the Decision Support System (DSS) perspective by demonstrating that

LLMs can function as decision-support mechanisms in investment decision-making. LLMs are capable of transforming heterogeneous market information into alternative investment signals in the form of Buy, Hold, and Sell recommendations. However, the findings from H1 indicate that these signals do not always correspond to actual market outcomes. Therefore, LLMs should be understood as augmentative technologies that enhance human decision-making capacity rather than replace human judgment.

Third, the findings contribute to the discussion surrounding the Efficient Market Hypothesis (EMH). Although significant differences were observed between LLM signals and actual market prices, the ability of all three models to generate positive cumulative returns suggests that AI-processed information may still possess economic value under certain market conditions. This finding indicates that market efficiency may not be absolute and that opportunities for extracting economically relevant signals may remain available in certain circumstances. Fourth, the findings are also relevant to Prospect Theory, particularly in understanding the relationship between investment return and risk. The differences between return performance and risk performance demonstrate that investors should not evaluate AI models solely based on their ability to generate high returns. DeepSeek generated the highest return, whereas ChatGPT exhibited the lowest investment risk. This finding highlights the importance of considering the trade-off between potential returns and potential losses when selecting AI-based investment-support tools. Conceptually, this study extends the literature on LLM applications in finance by demonstrating that AI effectiveness should not be evaluated solely through prediction accuracy. Instead, AI-based investment performance should be assessed multidimensionally through accuracy, profitability, and risk. Such an approach provides a more comprehensive framework for understanding the contribution of LLMs to investment decision-making.

C. Practical Implications

For investors, the findings indicate that ChatGPT, Gemini, and DeepSeek can be utilized as investment analysis tools, particularly for processing market information rapidly and systematically. However, AI-generated recommendations should not be used as the sole basis for investment transactions. Investors should combine AI recommendations with fundamental analysis, technical analysis, macroeconomic conditions, market sentiment, stock liquidity, and appropriate risk-management strategies. The differences in model performance also suggest that AI selection should be aligned with investors' objectives. If the primary objective is to maximize potential returns, DeepSeek demonstrated the strongest performance during the research period. Conversely, if investors prioritize risk mitigation, ChatGPT demonstrated relatively more conservative characteristics based on the risk indicators employed in this study. For the Financial Services Authority (OJK) and the Indonesia Stock Exchange (IDX), the increasing use of LLMs in investment activities highlights the importance of strengthening AI-assisted investment decision-making literacy. Investor education and regulatory frameworks should address AI transparency, recommendation errors, hallucination risks, and potential investor overreliance on automated recommendations. Retail investors should understand that AI-generated recommendations do not guarantee investment profits and cannot eliminate market risk.

For AI technology developers, the findings emphasize the importance of improving AI models' ability to understand the characteristics of the Indonesian capital market. Developing systems capable of integrating real-time market data, company fundamentals, technical indicators, local market sentiment, and macroeconomic information may improve recommendation quality. AI-based financial systems should also incorporate greater transparency so that users can understand the factors underlying specific recommendations. For educational institutions and financial practitioners, this study highlights the importance of developing financial literacy programs that integrate investment knowledge with AI capabilities. Such integration is necessary to ensure that users not only understand how to utilize AI technology but also possess the critical ability to evaluate the reliability and limitations of AI-generated recommendations.

D. Research Limitations

Despite providing empirical evidence regarding the application of LLMs in stock investment, this study has several limitations. First, the study employed a relatively short observation period, consisting of 20 trading days in December 2025. This period may not adequately capture model performance under different market regimes, including bullish, bearish, sideways, and crisis conditions. Second, the study focused exclusively on stocks included in the LQ45 Index. Because LQ45 stocks generally exhibit relatively high liquidity, the findings may differ when applied to small-capitalization stocks, low-liquidity stocks, or stocks characterized by higher volatility. Third, LLM performance is strongly influenced by prompt engineering. Although standardized prompts were employed to maintain experimental consistency, modifications in prompt structure, sequence, or information content may generate different recommendations. Therefore, the findings should be interpreted within the specific experimental prompt framework employed in this study.

Fourth, LLMs are dynamic technologies. Model updates, system modifications, parameter changes, and improvements in AI capabilities may cause the models to generate different outputs at different points in time even when identical prompts are used. This condition presents an important challenge for research replication. Fifth, the study did not comprehensively incorporate firm-level fundamental information. Variables such as earnings quality, valuation multiples, dividend policy, corporate governance, and financial health may provide additional information that could influence the quality of AI-generated investment recommendations. Sixth, the study primarily focused on accuracy, return, and investment risk and did not comprehensively examine investor psychology, behavioral biases, investor trust in AI, or actual investor behavior following AI-generated recommendations. These limitations indicate that the findings should be interpreted cautiously. Positive returns during the research period should not be interpreted as evidence that LLMs will consistently generate profits in future periods. Similarly, relatively low investment risk during the observation period does not imply that AI can eliminate investment risk.

E. Recommendations and Future Research Agenda

Based on the findings and limitations of this study, future research should employ longer observation periods to capture LLM performance across different market conditions. Multi-year datasets would enable researchers to assess the stability, robustness, and consistency of AI-generated recommendations across different market regimes. Future studies should also expand the research population beyond LQ45 stocks by including stocks with different levels of liquidity, capitalization, and volatility. Such expansion would improve the generalizability of findings concerning the effectiveness of LLMs in stock investment. Furthermore, future research should integrate fundamental analysis, technical indicators, social media sentiment, news sentiment, macroeconomic indicators, and real-time market data into experimental designs. Such an approach could determine whether greater information richness and higher-quality inputs improve the accuracy and profitability of LLM-generated recommendations. Future studies are also encouraged to employ more comprehensive performance measures, including the Sharpe Ratio, Sortino Ratio, Treynor Ratio, Jensen's Alpha, Value at Risk (VaR), Conditional Value at Risk (CVaR), maximum drawdown, and other risk-adjusted performance measures. These indicators would allow researchers to evaluate AI performance not only in terms of absolute returns but also in relation to the level of risk undertaken to generate those returns.

In addition, future research could compare a broader range of LLMs and machine-learning algorithms, including generative AI, predictive analytics, and hybrid AI models. Multi-model ensemble approaches could also be examined to determine whether combining recommendations from several AI models can produce more stable performance than relying on a single model. Finally, future research should develop approaches that integrate human intelligence and artificial intelligence. LLMs should not be conceptualized as substitutes for investors but as technologies

capable of enhancing human capacity to process information and evaluate investment alternatives. Future studies could therefore examine whether human–AI collaborative decision-making produces more accurate, profitable, and risk-efficient investment outcomes than decisions made exclusively by humans or AI systems. Overall, this future research agenda is expected to strengthen the empirical evidence concerning the role of LLMs in capital markets and contribute to the development of a more robust, transparent, adaptive, and responsible AI-assisted investment decision-making framework.

REFERENCE

Andersen, T. G., & Bollerslev, T. (2021). Volatility and trading volume: A persistent relationship. *Journal of Finance*, *76*(2), 1043–1091. https://doi.org/10.1111/jofi.12975

Borges, M. R. (2014). Efficiency testing of Asian stock markets. *International Journal of Finance \& Economics*, *19*(2), 140–155. https://doi.org/10.1002/ijfe.1489

Cardillo, A. (2025). *65 Most Popular AI Tools Ranked (October 2025)tle*. Exploding Topics.

Chawla, D., Bhutada, A., Anh, D. D., Raghunathan, A., SP, V., Guo, C., Liew, D. W., Gupta, P., Bhardwaj, R., Bhardwaj, R., & Poria, S. (2025). *Evaluating {AI} for Finance: Is {AI} Credible at Assessing Investment Risk Appetite?*

Chen J., W. Y., & Chen, X. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Finance Research Letters*, *56*, 103664. https://doi.org/10.1016/j.frl.2023.103664

Chen, J., Wang, Y., & Chen, X. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Finance Research Letters*, *56*, 103664. https://doi.org/10.1016/j.frl.2023.103664

Chen, X., Wang, J., & Zhong, X. (2024). Return Predictability of Prospect Theory: Evidence from the Thailand Stock Market. *Pacific-Basin Finance Journal*, *83*, 102199. https://doi.org/10.1016/j.pacfin.2023.102199

Chordia, T., Roll, R., & Subrahmanyam, A. (2020). Market liquidity and trading activity. *Journal of Financial Economics*, *135*(3), 644–670. https://doi.org/10.1016/j.jfineco.2019.06.008

Dowling, M., & Lucey, B. (2023a). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, *53*, 103662. https://doi.org/10.1016/j.frl.2023.103662

Eka, M. (2024). *Gemini AI :Kecerdasan Buatan Canggih Dari Google*. Direktorat Pusat Teknologi Informasi.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

Gartner. (2023). *Gartner Predicts More Than 80\% of Enterprises Will Use Generative {AI} {API}s or Deploy Generative {AI}-Enabled Applications by 2026*.

Hangyan. (2025). *印尼股票市场风险收益比 [{Rasio} risiko-return pasar saham {Indonesia}]*.

Jaglal, N. U., & Kose, U. (2024). Descriptive Statistics for Stock Risk Assessment. In *Data Analytics for Smart Cities*. CRC Press (Taylor & Francis).

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Ko, H., Lee, J., & Park, S. (2024). {ChatGPT} and portfolio diversification. *Finance Research Letters*, *59*, 104780. https://doi.org/10.1016/j.frl.2023.104780

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022a). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, *197*, 116659. https://doi.org/10.1016/j.eswa.2022.116659

Kustodian Sentral Efek Indonesia. (2025). *Jumlah investor pasar modal {RI} capai 20,32 juta {SID} di akhir 2025*.

Li, X., Zhang, R., & Zhou, N. (2022). A hybrid deep learning model integrating LSTM and Transformers for stock trend prediction. *Applied Intelligence*, *52*, 9800–9817. https://doi.org/10.1007/s10489-022-03371-1

Liu, Y., Jiang, Z., & Wang, H. (2023). Price--volume interaction and short-term stock return predictability. *International Review of Financial Analysis*, *86*, 102511. https://doi.org/10.1016/j.irfa.2023.102511

Marr, B. (2023). *A Short History Of ChatGPT: How We Got To Where We Are Today*. FORBES.

Marshall, B. R., Nguyen, N. H., & Visaltanachoti, N. (2021). Trading volume and technical analysis signals. *Journal of Banking \& Finance*, *122*, 105987. https://doi.org/10.1016/j.jbankfin.2020.105987

Martinez, R., Kumar, A., Singh, P., Chatterjee, S., Das, M., Roy, S., & Ghosh, D. (2026). Evaluating the efficacy of large language models in stock market decision-making: A decision-focused, price-only, multi-country analysis using historical price data. *Machine Learning and Knowledge Extraction*, *8*(4), 104. https://doi.org/10.3390/make8040104

Nazlioglu, S., Pazarci, S., Kara, A., & Varol, O. (2024). Efficient Market Hypothesis in Emerging Stock Markets: Gradual Shifts and Common Factors in Panel Data. *Applied Economics Letters*, *31*(18), 1773–1779. https://doi.org/10.1080/13504851.2023.2206613

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

Nison, S. (2001). *Japanese Candlestick Charting Techniques*. Penguin Publishing Group.

Oehler, A., & Horn, M. (2023b). Financial advice from robo-advisors versus large language models: Evidence from risk profiling and suitability. *Finance Research Letters*, *58*, 104222. https://doi.org/10.1016/j.frl.2023.104222

Otoritas Jasa Keuangan. (2025). *Tingkat melek \& akses pasar modal {RI} masih rendah, begini hasil surveinya*.

Pelster, M., & Val, J. (2024a). Can ChatGPT Assist in Picking Stocks? *Finance Research Letters*, *59*, 104786. https://doi.org/10.1016/j.frl.2023.104786

Pring, M. J. (2014). *Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points* (5th ed.). McGraw-Hill.

Rao, A., & Patel, M. (2023). Stock market prediction using GPT-based sentiment extraction: Evidence from high-frequency tweets. *Expert Systems with Applications*, *228*, 120310. https://doi.org/10.1016/j.eswa.2023.120310

Schneider, J., & Yilmaz, F. (2025). Multifaceted variability in {LLM}-driven stock recommendations. *Journal of Banking \& Finance*, *169*, 107320. https://doi.org/10.1016/j.jbankfin.2025.107320

Setiawati, S. (2025). *Sosok di Balik DeepSeek yang Bikin AI Amerika Ketar-Ketir*. CNN Indonesia.

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, *90*, 106181. https://doi.org/10.1016/j.asoc.2020.106181

Simon, H. A. (1957). *Models of Man: Social and Rational*. John Wiley \& Sons.

Solares, E. A., Aiello, L. M., & others. (2022). A comprehensive decision support system for stock market investment. *Decision Analytics Journal*, *4*, 100075. https://doi.org/10.1016/j.dajour.2022.100075

Technology, T. (2025). *Top 10 Most Used Gen AI Tools In 2025*. Titan Technology.

Unesa. (2025). *DeepSeek: Terobosan AI Hemat Biaya yang Mengguncang Dunia Teknologi*. Fakultas Teknik Universitas Negeri Suraba.

Xu, L., Hu, Z., & Li, C. (2024). Large language models and market efficiency: Evidence from news-based trading signals. *Journal of Financial Markets*, *67*, 100843. https://doi.org/10.1016/j.finmar.2023.100843

Zhang, Y., & Chen, W. (2022). Volume-confirmed breakout strategies and stock market performance. *Finance Research Letters*, *47*, 102785. https://doi.org/10.1016/j.frl.2021.102785

Andersen, T. G., & Bollerslev, T. (2021). Volatility and trading volume: A persistent relationship. *Journal of Finance*, *76*(2), 1043–1091. https://doi.org/10.1111/jofi.12975

Borges, M. R. (2014). Efficiency testing of Asian stock markets. *International Journal of Finance \& Economics*, *19*(2), 140–155. https://doi.org/10.1002/ijfe.1489

Cardillo, A. (2025). *65 Most Popular AI Tools Ranked (October 2025)tle*. Exploding Topics.

Chawla, D., Bhutada, A., Anh, D. D., Raghunathan, A., SP, V., Guo, C., Liew, D. W., Gupta, P., Bhardwaj, R., Bhardwaj, R., & Poria, S. (2025). *Evaluating {AI} for Finance: Is {AI} Credible at Assessing Investment Risk Appetite?*

Chen, J., Wang, Y., & Chen, X. (2023). Can ChatGPT forecast stock price movements? Return

predictability and large language models. *Finance Research Letters*, *56*, 103664. https://doi.org/10.1016/j.frl.2023.103664

Chen, X., Wang, J., & Zhong, X. (2024). Return Predictability of Prospect Theory: Evidence from the Thailand Stock Market. *Pacific-Basin Finance Journal*, *83*, 102199. https://doi.org/10.1016/j.pacfin.2023.102199

Chordia, T., Roll, R., & Subrahmanyam, A. (2020). Market liquidity and trading activity. *Journal of Financial Economics*, *135*(3), 644–670. https://doi.org/10.1016/j.jfineco.2019.06.008

Dowling, M., & Lucey, B. (2023a). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, *53*, 103662. https://doi.org/10.1016/j.frl.2023.103662

Eka, M. (2024). *Gemini AI :Kecerdasan Buatan Canggih Dari Google*. Direktorat Pusat Teknologi Informasi.

Gartner. (2023). *Gartner Predicts More Than 80\% of Enterprises Will Use Generative {AI} {API}s or Deploy Generative {AI}-Enabled Applications by 2026*.

Jaglal, N. U., & Kose, U. (2024). Descriptive Statistics for Stock Risk Assessment. In *Data Analytics for Smart Cities*. CRC Press (Taylor & Francis).

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Ko, H., Lee, J., & Park, S. (2024). {ChatGPT} and portfolio diversification. *Finance Research Letters*, *59*, 104780. https://doi.org/10.1016/j.frl.2023.104780

Marr, B. (2023). *A Short History Of ChatGPT: How We Got To Where We Are Today*. FORBES.

Nazlioglu, S., Pazarci, S., Kara, A., & Varol, O. (2024). Efficient Market Hypothesis in Emerging Stock Markets: Gradual Shifts and Common Factors in Panel Data. *Applied Economics Letters*, *31*(18), 1773–1779. https://doi.org/10.1080/13504851.2023.2206613

Pelster, M., & Val, J. (2024c). Can large language models assist in picking stocks? *Finance Research Letters*, *59–60*, 104292. https://doi.org/10.1016/j.frl.2023.104292

Pring, M. J. (2014). *Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points* (5th ed.). McGraw-Hill.

Schneider, J., & Yilmaz, F. (2025). Multifaceted variability in {LLM}-driven stock recommendations. *Journal of Banking \& Finance*, *169*, 107320. https://doi.org/10.1016/j.jbankfin.2025.107320