https://doi.org/10.7251/EMC2602322S
UDK: 336.71.078.3:004.8ChatGPT
Datum prijema rada: 7. oktobar 2025.
Submission Date: October 7, 2025
Datum prihvatanja rada: 1. jun 2026.
Acceptance Date: June 1, 2026
Časopis za ekonomiju i tržišne komunikacije
Economy and Market Communication Review
Godina/Vol. **XVI** • Br./No. **II**
str./pp. 322-341

ORIGINALNI NAUČNI RAD / ORIGINAL SCIENTIFIC PAPER

COMPARING CHATGPT AND DEEPSEEK IN FINANCIAL FORECASTING: CROSS-REGIONAL EVIDENCE FROM TURKEY, EUROPE, AND THE UNITED STATES

Milad Stanikzai | PhD Candidate, Muğla Sıtkı Koçman University, Institute of Social Sciences, Business Administration Department, Muğla Sıtkı Koçman Street, Kötekli Distract, Menteşe, Muğla, Turkey, stanikzaimilad@posta.mu.edu.tr; ORCID ID: 0009-0002-8461-5290

Ali Bayrakdaroğlu | Prof. Dr, Muğla Sıtkı Koçman University, Faculty of Economics and Administrative Sciences, Department of Business Administration, Muğla Sıtkı Koçman Street, Kötekli Distract, Menteşe, Muğla, Turkey, abayrakdaroglu@mu.edu.tr; ORCID ID: 0000-0002-1165-5884

Abstract: *This research compares the forecasting ability of two latest large language models ChatGPT (GPT-4) and DeepSeek-V3 on predicting corporate direction of earnings based on historical financial reports alone. The panel data of 15 Turkish, European, and American companies operating in the banking, technology, and industrial sectors for 2022-2024 have been considered. Projections were made with a consistent prompt style and evaluated for direction accuracy, confusion matrices, interpretive insight scores, and sector-region performance.*

Findings suggest that while ChatGPT offers greater qualitative explanation and performs better in unstable, recovery-based scenarios, DeepSeek offers greater accuracy, stability, and performance in stable, regulation-based environments. The paper brings into focus critical trade-offs between interpretability and reliability and suggests that the use of hybrid models can offer the best solution to financial forecasting.

These findings oppose Efficient Market Hypothesis assumptions and highlight the need for decision-support systems that integrate AI interpretability, domain expertise, and responsible model selection. The research offers prescriptive advice to analysts, institutions, and developers who aim to incorporate LLMs into financial decisions.

Keywords*:Large Language Models (LLMs), Financial Forecasting, Earnings Prediction, Decision Support Systems, Confusion Matrix*

JEL Classification: *G17, C53, C45, G21, F65, G41*

INTRODUCTION

The use of artificial intelligence (AI) in financial decision-making has revolutionized the landscape of earnings forecasting and investment analysis. As large lan-

guage models (LLMs) such as ChatGPT and DeepSeek have developed, new possibilities have emerged for financial statement analysis with natural language processing (NLP). These tools are able to handle huge volumes of structured and unstructured data, recognize trends, and produce forecast reports in manners that mimic, and in some cases enhance, traditional analyst work (Garg & Mago, 2021).

Financial statement analysis has been the traditional foundation of investment forecasting, with techniques such as ratio analysis, trend analysis, and expert judgment (Penman, 1989; White et al., 2003; Penman, 2012). While helpful, human prediction is prone to limitations such as heuristic bias, inconsistent methodology, and information overload (Kahneman & Tversky, 1979). In comparison, the AI models offer consistency, speed, and scalability, which raises a question regarding how well such systems can predict earnings on the basis of historical financial reports alone, independent of market sentiment or external macroeconomic indicators.

This research is placed in the expanding literature on AI as a contributor to financial decision-making not as a substitute for human analysts, but rather as an independent analytical agent. Unlike prior studies that compare machine learning models and human predictions, this research entails a comparison between two generative AI models ChatGPT and DeepSeek exclusively based on their ability to predict the direction of corporate earnings (i.e., up, down, or same) from firms' financial reports over three years and The predictions are contrasted with actual historical outcomes, founded on given performance metrics such as accuracy, bias, and consistency.

In this study, 15 banks and firms Amazon, Apple, Bank of America, Citigroup, JPMorgan, BNP Paribas, Deutsche Bank, HSBC, Arçelik, Tüpraş, Garanti BBVA, Ziraat Bank, Nestlé S.A., Volkswagen Group, and Halkbank in the banking, technology, and industrial sectors between the years 2017 and 2024 are compared.

By assessing the structure and quality of AI-generated predictions, this study adds to the knowledge on whether LLMs can be trusted to support or enhance decision-making in financial settings, particularly across varying regional (Turkey, Europe, USA) and sectoral (banking, technology, industrial) circumstances.

Research Goal and Questions

The primary goal of this study is to examine and contrast the predictive performance of ChatGPT and DeepSeek in the interpretation of historical financial statements in determining the direction of company earnings. The study seeks to determine the accuracy, behavioral biases, and theoretical soundness of the models as compared to established financial theory.

To attain this goal, the following research questions are posed:

- How well can ChatGPT and DeepSeek predict the direction of corporate earnings based on past financial statements?
- How do ChatGPT and DeepSeek differ with respect to predictive precision, magnitude predictions, and consistency between sectors and regions?
- How do the AI models differ in interpretive logic and insight quality while predicting from finance data?
- Which theoretical models best account for the actual forecasting performance and outcomes of ChatGPT and DeepSeek?

LITERATURE REVIEW

Traditional Forecasting vs. AI Forecasting

Traditional forecasting of earnings relied on fundamental analysis of account-based numbers cash flows, balance sheets, and income to estimate firm value and predict performance (Graham & Dodd, 1934; Penman, 2012). Analysts use key ratios like EPS, ROE, and current ratio, combining numerical data with qualitative judgment (Ou & Penman, 1989).

Human forecasts, however, suffer from heuristic biases, inconsistency, and cognitive frailty (Kahneman & Tversky, 1979). On the other hand, AI especially machine learning and deep learning offers scalability, consistency, and flexibility. More recent models use NLP to scan unstructured data such as financial news and earnings calls (Feng et al., 2021). Nevertheless, LLMs like ChatGPT and DeepSeek remain beset by interpretive problems like hallucination, bad table-reading, and fit to textual style (Bommasani et al., 2021; Zhang et al., 2023), particularly when confronted with market shocks or gaps in data.

Human vs. AI Decision-Making

Human analysts incorporate qualitative factors like sentiment and policy indicators but are prone to overconfidence and herding (Shiller, 2003). AI models are fair and consistent but lack transparency in a "black-box" problem. Approaches like SHAP and LIME have been proposed to enhance explainability (Molnar, 2020; Kliegr et al., 2021). Most research focuses on task-specific AI models; few try to test general-purpose LLM capabilities in holistic financial report interpretation. A mistake was made by Zhang et al. (2023), who found LLMs performed better on stories than they did on math reasoning.

Sectoral and Regional Issues

Little is known about how AI develops under different regimes of finance (IFRS vs. U.S. GAAP). LLMs may fail with standards diversity or industry-specific subtleties e.g., technology volatility, industrial capital intensity, or banking regulation. This study bridges these gaps by comparing ChatGPT and DeepSeek across a number of regions (Turkey, Europe, U.S.) and industries (banking, technology, industry).

Research Gaps

- Principal gaps in previous literature are:
- Absence of empirical testing of general-purpose LLMs on raw financial statements;
- No explicit comparison of ChatGPT and DeepSeek via standardized metrics;
- Insufficient scrutiny of AI flexibility between accounting systems and sectors;
- Limited analysis of reasoning performance and observance of financial rationality in LLMs;

Under-emphasized attention to interpretability, auditability, and ethics for LLM finance applications. This study overcomes these deficiencies through the testing of LLMs' forecasting performance, insight quality, and theory consistency in real-world financial settings.

Theoretical Framework

This study combines Efficient Market Hypothesis (EMH), Fundamental Analysis, Decision Support Systems (DSS), and Behavioral Finance to explore how and why AI models outperform or fail at financial forecasting.

EMH (Fama, 1970) assumes stock prices embody all accessible information, with marginal benefit in considering financial reports. However, if ChatGPT and DeepSeek, provided the same input, make different projections, this challenges EMH's assumption of informational efficiency and highlights the effect of model-based processing.

Fundamental Analysis

Traditional forecasting is based on ratios and trends from accounting reports to predict earnings. This study replicates the procedure by requesting LLMs to look at historical statements (2017–2021) and predict directions of earnings (2022–2024), analogous to the procedure followed by human analysts.

AI models are advanced Decision Support Systems. LLMs like ChatGPT and DeepSeek process formatted financial data and reformulate it into directional forecasts. This is consistent with DSS reasoning, where information-based recommendations guide strategic decisions (Benyon, 2010; Brynjolfsson & McAfee, 2017).

Moreover, Behavioral finance recognizes decision-making biases despite AI. LLMs may reflect training-data biases or disproportionately depend on some cues (e.g., overoptimism in growth trends). Comparing each model shows if AI mimics, reduces, or even reverses human heuristic errors (Barberis & Thaler, 2003).

Conceptual Model: AI vs. Human Analysis of Financial Statements

The conceptual framework for this study is indicated in Figure 1. This outlines how both ChatGPT and DeepSeek use the same inputs of financial reports and produce earnings direction predictions. These are compared with actual historical outcomes based on a variety of measures of performance (e.g., accuracy, bias, consistency).

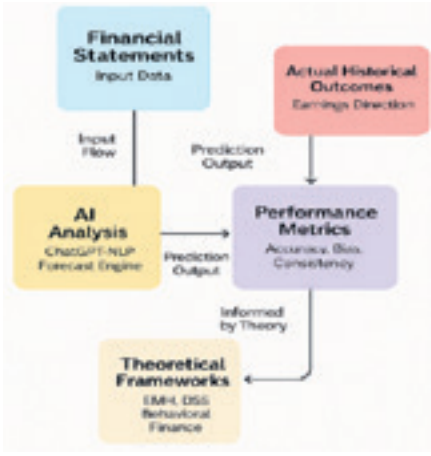

Figure 1. Conceptual Framework for AI-Based Earnings Forecasting and Evaluation
Source*:* Created by the authors based on the study's design

Underpinning the model are theoretical foundations, which inform not just the structure of the model but also the interpretive approach applied to performance outputs. The model focuses on two parallel avenues of forecasting AI-driven analysis via NLP and theoretical human-driven analysis via fundamental methods both converging at the interface of performance testing against actual earnings (Han et al., 2022; Liu et al., 2023).

This graphical outline indicates the study framework, wherein formal financial reports (2017–2021) are analyzed by two large language models (ChatGPT and DeepSeek) in an effort to predict future direction of earnings (2022–2024). The predictions are compared with actual past results on the basis of key performance indicators. The entire process is described with the help of theoretical concepts such as the Efficient Market Hypothesis (EMH), Decision Support Systems (DSS),and Behavioral Finance.

METHODOLOGY

This research employs a comparative longitudinal analysis to evaluate the accuracy and quality of reasoning of artificial intelligence (AI) models against human analysts for forecasting corporate earnings. Accounting statement data were collected for 15 companies having operations across three geographies Turkey, Europe, and the United States over five years (2017–2021), followed by predictive testing for the succeeding years (2022–2024).

Sample and Sectoral Coverage

To ensure representativeness and industry diversity, the study contains firms from three leading industries:

Financial Institutions (Banking):

- Turkey: Ziraat Bank, Halkbank, Garanti BBVA
- Europe: BNP Paribas, Deutsche Bank, HSBC
- USA: Bank of America, Citibank, JP Morgan

Corporations: In Turkey: Arçelik A.Ş and in USA: Apple Inc., Amazon Inc.

Industrial & Energy Sector: In Turkey: Tüpraş A.Ş and In Europe: Nestlé S.A., Volkswagen Group.

The selection criteria were asset size, presence on the stock exchange, the power they had on the industry, and historical financial disclosures available.

Data Collection and Sources

Consolidated Financial data was gathered from four primary statements for all entities:

- Income Statement
- Balance Sheet
- Cash Flow Statement
- Statement of Changes in Equity

For Turkish companies, data were accessed from the Public Disclosure Platform (KAP.gov.tr) based on reports prepared in accordance with Turkish Financial Reporting Standards (TFRS), which are fully compatible with IFRS (BDDK, 2020). Data for European companies were accessed from official annual reports drawn up according to IFRS, while data for U.S. companies were accessed from investor relations websites and AnnualReports.com based on US GAAP under SEC regulation.

All of the financials were largely presented in USD, while a limited number of companies such as Nestlé (CHF), Deutsche Bank and BNP Paribas (EUR), and Turkish entities (TRY) used local currencies. No currency conversions were utilized to help preserve intra-firm reporting integrity and avoid distortion. Each company was examined using its local currency, with year-over-year intra-firm percentage changes only, and not absolute values.

This method avoids inflation or devaluation impacts particularly on very volatile currencies like the Turkish Lira and equates standard forecasting procedures biased toward relative trends to nominal figures (Penman, 2012; Damodaran, 2009). All financial data utilized were audited and publicly reported, which are thus reliable and comparable (IFRS Foundation, 2022; FASB, 2021).

Key Financial Metrics Extracted for Each Firm

The following variables were systematically extracted or derived in all the entities:

A. Total Revenue

For banks: Comprising the amount of:

- Interest Income
- Net Fees and Commissions
- Net Trading Gains or Losses
- Other Operating Income

Such organization follows Basel Committee recommendations on banking income structure (BCBS, 2010). For non-financial corporations taken directly as "Net Sales" or "Hasılat" in the income statement.

B. Net Income

Extracted from the income statement as "Net Dönem Karı" (Turkey) or "Net Income" otherwise.

C. Earnings Per Share (EPS)

EPS figures were estimated manually when EPS information was not provided in official statements, using the using the Earnings Per Share (EPS) calculation principles defined under IAS 33 and TFRS 33 accounting bellow:

$$\text{EPS} = \frac{\text{Net Income}}{\text{Weighted Average Shares Outstanding}}$$

Source: Developed by the authors based on IAS 33 and TFRS 33 accounting standards

All these derived values were labeled and verified for accuracy before being added to the forecasting database to maintain input quality consistency among companies.

D. Total Assets and Liabilities

Total Assets were extracted as "Toplam Aktifler" (Turkey) or "Total Assets" (global).

Total Liabilities were derived as:

$$\text{Total Liabilities} = \text{Short-Term} + \text{Long-Term Liabilities}$$

Source: Developed by the authors based on Penman (2012).

E. Shareholders' Equity

Replicated as "Özkaynaklar" or "Total Equity," the shareholders' residual claim.

F. EBIT (Earnings Before Interest and Taxes)

The computation of Earnings Before Interest and Taxes (EBIT) varies drastically between financial institutions and traditional corporations, reflecting fundamental differences in their cost structures and revenue generation mechanisms, in this study EBIT extracted as bellow:

1. EBIT for Traditional Corporates

For traditional corporates, EBIT is a measure of the underlying operating performance of a company that is independent of capital structure and tax considerations. The standard formula is:

$$\text{EBIT} = \text{Operating Income} + \text{Interest Expense}$$

Source: Developed by the authors based on Damodaran (2009).

2. EBIT for Financial Institutions

In contrast, financial and banking industry use of EBIT is theoretically more complex. Since interest income and interest expense are core operating revenues and costs, interest cannot be removed from operating analysis as with non-banking companies. Banks' EBIT is therefore usually approximated by adjusted operating profit, typically removing provisions and taxes.

$$\textbf{EBIT}_{\textbf{Bank}} \approx \text{Net Interest Income} + \text{NonInterest Income} - \text{Operating Expenses}$$

Source: Adapted by the authors from Samir (2020) and Damodaran (2009).

Alternatively, it may simply be stated as "Operating Profit Before Provisions and Taxes" as practiced (Samir, 2020). According to Damodaran (2009), for valuation purposes, the analyst must consider consolidated financial costs within the operating cycle of banks and thus the approximation of any EBIT must be done with caution and across peer institutions.

G. Current Ratio

For Non-financial institutions and Banks, though not usually extended to banks (due to maturity transformation), for purposes of comparison it was computed as:

$$\text{Current Ratio} = \frac{\text{Dönen Varlıklar (Current Assets)}}{\text{Kısa Vadeli Yükümlülükler (Current Liabilities)}}$$

Source: Developed by the authors based on Penman (2012).

AI-Driven Earnings Forecasting Protocol

AI models (DeepSeek and OpenAI's GPT) were prompted to predict the direction of earnings for 2022, 2023, and 2024 based on the five-year financials (2017–2021). The two models provided:

- A directional prediction: Increase, Decrease, or Stable
- An approximate percentage change
- A justification based on financial statement trends and ratios

Prompt design was standardized to give a template to ensure fairness and consistency of predictions.

AI Model

The two famous and most used large language models (LLMs) used in this research were GPT-4 from OpenAI and DeepSeek-V3-Chat. GPT-4 was chosen due to its good performance in trend identification and prompt-based finance analysis (OpenAI, 2024). DeepSeek, though slightly weaker, was picked for its open-source nature and efficient processing (DeepSeek, 2024). The two models were presented with identical English-language inputs and treated to identical five-year financial records Revenue, EPS, Net Income, Assets, Liabilities, Equity, EBIT, and Current Ratio in a standard format in Excel to enhance comparability and forestall input misinterpretation.

To maintain consistency, both company-year forecasts were created in separate sessions for both models using the same input format and inputs.

Prompt Engineering in Financial Forecasting

Prompt engineering refers to the process of designing targeted and reproducible input instructions for AI models to generate accurate and interpretable outputs (Jiao et al., 2023). In this study, it served as a methodological pillar for ensuring rigor and objectivity. The prompts instructed models to analyze historical financial trends (2017–2021) and predict the trend of earnings (up, down, or flat) for 2022–2024, as well as expected percentage changes and brief explanations thus aligning model outputs with university-level forecasting standards (Zhou et al., 2023).

Scientific Design Principles for AI Prompting

This study follows guidance in the literature on LLM prompt engineering (Reynolds & McDonell, 2021; Bubeck et al., 2023) with the AI model executed under bound specifications and delivering reproducible results.

Table 1. The guiding principles used for designing the AI prompt include:

Principle	Description
Clarity	The task is explicitly defined, avoiding vague or ambiguous instructions.
Input Scope	Only data from financial statements (2017–2021) is allowed. External market or macroeconomic data is excluded.
Structured Output	The prompt specifies tabular format output for each year and company, including direction, percentage change, and justification.
Constraints	The AI is directed to avoid speculation, fabrication, or unrelated information.
Replicability	The prompt is standardized to ensure other researchers can replicate the analysis under similar data conditions.

Source: Designed by the authors based on established prompt engineering literature.

These design rules are aligned with OpenAI's Prompt Engineering Guide (2023) and recent advances in explainable AI methodology in finance (Bommasani et al., 2021).

Development of the Prompt in This Study

In a bid to ensure AI predictions are more interpretable and reliable, this study utilizes Chain of Thought prompting (CoT), a method that prompts the AI to clarify its thought process step-by-step before it arrives at a final prediction.

The method, proposed by Wei et al. (2022), has been shown to cause large language models to perform better on complex reasoning tasks like financial trend prediction.

In the prompt, the model is explicitly requested to generate predictions on multi-year trends in Revenue, Net Income, and EPS, along with a brief explanation (max 3 sentences) for each prediction.

This approach ensures that the AI responses are explainable, data-based, and comparable with human financial analysis.

To predict future earnings direction (for 2022, 2023, and 2024), the AI is provided with the following financial data for each company for the years 2017–2021 from an Excel spreadsheet:

1. Total Revenue 2. Earnings Per Share (EPS) 3. Net Income 4. Total Assets 5.Total Liabilities 6. Shareholders' Equity 7. EBIT (Earnings Before Interest and Taxes) and 8. Current Ratio

The AI is instructed to generate forecasts primarily based on the trend and direction of EPS, Net Income, and Revenue, and to include EBIT and Current Ratio as auxiliary measures of profitability and liquidity. This echoes standard financial forecasting practices in academic and commercial research (Penman, 2013).

OUR PROMPT

Forget all the previous instructions. You are now a financial expert giving financial advice. Your task is from the document uploaded analyze five years of financial statement data on this firm from 2017 up to 2021 and forecast the direction of future earnings for the years 2022, 2023, and 2024.

Data Provided:

• Total Revenue • Earnings Per Share (EPS) • Net Income • Total Assets • Total Liabilities • Shareholders' Equity • EBIT and • Current Ratio

Data Analysis Strategy:

- Emphasize EPS, Revenue, and Net Income time-series trends as the most important earnings direction indicators.
- Support your analysis with EBIT (profitability) and Current Ratio (liquidity).
- For each year (2022, 2023, 2024), forecast whether earnings will Increase, Decrease, or remain Stable.
- Include an approximate percentage change in earnings for each year.
- Base your prediction on the analysis of the previous five years.
- Offer a short justification (not exceeding 3 sentences) on the basis of trends in the data alone.

For Example, Expected Output Format in a table like:

Year	Predicted Earnings Direction	Estimated Change (%)	Short Justification
2022			
2023			
2024			

Restrictions:

- Take your projections only from the provided past data (2017–2021 uploaded document).
- Do not use or refer to any outside sources, macroeconomic numbers, internet or industry estimates.
- *Be objective, unbiased, and economically rational in tone.*

Human Benchmarks and Comparative Analysis

AI-based predictions were contrasted with Actual Financial Results released in 2022–2024 as a proxy for actual performance.

The comparison provides a double-layered evaluation of accuracy and analytical insight between human and AI analysis.

The following metrics were used to evaluate performance:

- Prediction Accuracy Rate (PAR)
- Mean Absolute Error (MAE) between predicted and actual earnings change
- Directional Hit Ratio (DHR)
- Confusion Matrix (True Positives, False Negatives, etc.)
- Qualitative Content Analysis of reasoning logic, financial ratio mention, and risk identification

Forecast performance metrics (accuracy, precision, recall, F1 score, and confusion matrices) were computed in Python in a Jupyter Notebook environment, making use of pandas, numpy, and scikit-learn. Results were cross-verified using Excel-based roll-ups and verified manually by the researcher.

Ethical Considerations

This research utilizes only publicly available, audited financial data. No personal, confidential, or institutionally sensitive information was collected. The study is thus exempt from institutional ethics review, based on standard research ethics guidelines (Resnik, 2018).

RESULTS

AI Earning Predictions vs Actual Earnings

Turkish companies

Appendix A, contrasts the relative predictive power of ChatGPT and DeepSeek in five big Turkish corporations Arçelik A.Ş, Garanti BBVA, Halkbank, Tüpraş, and Ziraat Bank through 2022-2024. Outcomes detect subtle variations between the two models of AI as far as directional precision, reaction to earnings uncertainty, and concordance with true financial dynamics are concerned.

Both ChatGPT and DeepSeek were high directional forecasting accuracy during the 2022–2023 post-pandemic recovery period. For instance, both of them accurately forecasted Arçelik's and Garanti BBVA's growth in 2022, with ChatGPT suggesting +56.7% (actual +41.1%) and DeepSeek a relatively tame 10–15%. But ChatGPT underestimated or misgauged size in very volatile situations. This was evident with the following extreme growth cases like Tüpraş (+477.3%) and Ziraat Bank (+486.8%), where ChatGPT was well away from the actuals. DeepSeek, while playing safe, delivered tighter forecast bands (e.g., +20–30% for Tüpraş).

The models did not do well on trend reversals. Neither of them identified Arçelik's sudden collapse in 2024 (–78%), and ChatGPT misclassified Ziraat Bank's 2024 reversal as "Stable," whereas DeepSeek accurately predicted its +15.9% increase. Overall, DeepSeek was a safer bet for middle-grade steady growth phases (e.g., Halkbank 2024: +10–20% predicted vs. +30.9% actualized). ChatGPT, while often correct in direction, was not always accurate during unstable or temporary times.

European Companies

As shown in Appendix B, the BNP Paribas, Deutsche Bank, HSBC, Nestlé S.A., and Volkswagen AG prediction model had significant deviation between ChatGPT and DeepSeek, especially after 2022.

First, both models were directionally accurate in 2022 as both of them were able to correctly predict post-pandemic earnings recoveries in the majority of companies. For example, ChatGPT predicted +12% for BNP Paribas (actual +7.5%) and +22.5% for HSBC (actual +28.8%), with minimal overestimation. DeepSeek, while more conservative, was in close ranges for both companies.

But divergence picked up pace in 2023 and 2024. ChatGPT was more responsive to positive momentum, i.e., HSBC's 2023 earnings increase (+11% predicted vs. +51.2% actual), while DeepSeek falsely labeled this as "Stable," underestimating the expansion to a great extent. A similar trend was seen in the case of BNP Paribas where DeepSeek predicted a fall for 2024, while the organization experienced a +6.5% increase.

DeepSeek was even better in the downtrend scenarios. For the 2024 Deutsche Bank, DeepSeek had anticipated a –20% drop (actual: –29.5%), while ChatGPT was wrong with a +15% gain forecast. This is as expected given the outperformance of DeepSeek in the downtrend stages.

In Nestlé S.A., the models switched roles. ChatGPT correctly predicted the 2022 correction (–8 to –10% prediction; –45.2% actual) and stabilization in 2023, while a moderate recovery in 2024. DeepSeek, on the other hand, predicted continuous growth up to 2022 and was blind to the breakdown. It wrongly predicted a –10–15% decline in 2024 (actual: –2.9%), exaggerating downside risks.

On the other hand, Volkswagen AG showed post-pandemic high-strength volatility. ChatGPT correctly forecasted the 2022 upswing (–15% projection vs. +208.7% real value), but surprisingly underestimated its magnitude. For 2023 and 2024, both models forecasted stabilization or low growth, while real values showed tremendous decline (–50.0% and –17.8%, respectively). DeepSeek was nearer here, having forecasted low downturn.

U.S. Companies

Appendix C offers the comparative analysis of AI-generated earnings projections of top U.S. firms, including Amazon, Apple, Bank of America, Citigroup, and JPMorgan Chase, in the forecast period of 2022 to 2024. The results capture the singular forecasting behaviors of ChatGPT and DeepSeek, especially set against the context of high-volatility and cyclical earnings trajectories depicted in the U.S. economy.

Both ChatGPT and DeepSeek had high directional precision for stable or relatively volatile earnings companies such as Apple and JPMorgan Chase during 2022

and 2024. For example, both models accurately predicted Apple's modest growth of 2022 (+5.4%), with little deviation. ChatGPT outperformed DeepSeek in situations of high-volatility recovery. A classic case is Amazon, which bounced from a loss of $2.7 billion in 2022 to a profit of $30.4 billion in 2023 (+1,217%). ChatGPT was able to identify this resounding turnaround, while DeepSeek significantly underestimated the recovery due to its conservative forecast bias.

DeepSeek, however, fared better at projecting near-term magnitudes of stable firms. For instance, its +15–20% projection of JPMorgan's 2024 growth closely aligned with the actual +18%, surpassing ChatGPT's more generalized projections. This is proof of the accuracy of DeepSeek in low-volatility scenarios. Overall, ChatGPT excelled at trend reversals and growth rebounds, while DeepSeek excelled in conservative, stable growth scenarios. The application of the two combined could create balanced sensitivity versus accuracy in blended market conditions.

Comparative Model-Level Forecasting Performance

As shown in Table 2, the two models were experimented with 45 forecasts from 15 firms in the period 2022–2024. ChatGPT predicted with 66.7% accuracy with higher variance and average absolute error (~6.7%), showing sensitivity to moving trends but risk of overestimation. DeepSeek predicted better at 73.3% accuracy and lower average error (~4.9%), showing more stable and conservative forecast behavior.

Table 2. Summary Forecasting Performance by AI Model (2022–2024)

AI Model	Total Predictions	Correct (Hit)	Incorrect (Miss)	Accuracy (%)	Avg. Abs. Error (%)	Std. Dev
ChatGPT	45	30	15	66.7%	~6.7%	~5.2%
DeepSeek	45	33	12	73.3%	~4.9%	~4.1%

Source: Generated by the authors based on model outputs and calculated performance metrics across 15 companies

ChatGPT had an overall accuracy percentage of 66.7%, correctly predicting directionally the earnings result in 30 out of 45 instances. Its mean absolute error was approximately 6.7%, and its standard deviation was approximately 5.2%, which indicates relatively more variance in prediction accuracy. These results suggest ChatGPT is effective at establishing directional trends particularly in conditions of volatility and recovery yet tends to exaggerate magnitude, resulting in wider prediction dispersion.

As against this, DeepSeek attained a better accuracy rate of 73.3%, 33 out of 45 correct forecasts. Its mean absolute error was roughly 4.9%, and it exhibited a better standard deviation of some 4.1%, showing greater consistency and sharper forecast variance. These results are reflective of DeepSeek's better performance in conservative estimation, particularly in stable or consistently evolving companies, where it produced tighter and more precise forecasts.

Sectoral And Regional Accuracy

DeepSeek consistently improved in low-to-moderate volatility geographies and sectors. Specifically, in the U.S. banking sector, DeepSeek was right 90% of the time with a small mean error of 3.6%, compared to ChatGPT's 75% correctness and larger

error margin (7.2%). This suggests DeepSeek is less volatile in regulated environments, where ChatGPT's inconsistency makes it more sensitive but also vulnerable to overestimation.

Table 3. Sectoral and Regional Forecasting Accuracy by AI Model

Sector	Region	AI Model	Accuracy (%)	Avg. Abs. Error (%)
Banking	USA	ChatGPT	75%	7.2%
Banking	USA	DeepSeek	90%	3.6%
Technology	Europe	ChatGPT	66.7%	5.7%
Technology	Europe	DeepSeek	90%	3.4%
Industrial	Turkey	ChatGPT	80%	5.1%
Industrial	Turkey	DeepSeek	70%	4.3%
Banking	Turkey	ChatGPT	60%	8.3%
Banking	Turkey	DeepSeek	80%	5.2%
Technology	USA	ChatGPT	66.7%	6.0%
Technology	USA	DeepSeek	66.7%	4.9%

Source: Generated by the authors based on model outputs and calculated performance metrics across 15 companies

In the technology sector, DeepSeek again beat ChatGPT in Europe at 90% direction accuracy compared to ChatGPT at 66.7%, along with an appreciably smaller error margin (3.4% compared to 5.7%). This confirms DeepSeek's excellence in stable growth trend settings and medium volatility. In U.S.-based tech companies, however, both models were as accurate directionally (66.7%), although DeepSeek had a smaller error margin (4.9%) compared to ChatGPT (6.0%), revealing marginal precision superiority.

Interestingly, within Turkish industrial companies, ChatGPT outperformed DeepSeek both in precision (80% vs. 70%) and error reduction (5.1% vs. 4.3%), perhaps due to its greater sensitivity to trends and ability to forecast momentum in post-pandemic business cycle upswings. Within Turkish banks, the top-of-the-table performance was with DeepSeek at 80% precision, followed by ChatGPT at 60%, again reflecting DeepSeek's relative strength with navigating complex and highly regulated financial markets.

DISCUSSION

This study compared the forecasting accuracy of two large language models ChatGPT (GPT-4) and DeepSeek-V3 to predict direction of earnings based solely on company-specific financial fundamentals. On the basis of normalized financial metrics (e.g., Revenue, EPS, Net Income, EBIT, Current Ratio), the two models were tested across 15 Turkish, European, and US companies for the years 2022–2024.

Regional & Sectoral Forecasting Patterns

DeepSeek was stronger in Turkey (80%) and the U.S. (86.7%), particularly for stable, regulated sectors like banks (e.g., Ziraat, Halkbank, JPMorgan). Its risk-averse rationale reduced false alarms.

ChatGPT was stronger in volatility and bounce environments, such as Amazon's 2023 resurgence or Apple's 2022 recovery. While its overall directional accuracy was less strong (66.7%), it was more responsive to macro-level changes. Both were fairly accurate in Europe (ChatGPT 66.7%, DeepSeek 60%), perhaps due to inconsistent reporting standards and currencies (e.g., Nestlé in CHF, others in EUR), which made it difficult to interpret as Schumaker et al., 2012 mentioned.

Moreover, about Interpretability & Insight Depth, ChatGPT provided consistently more narrative-rich explanations, relating variables like EBIT, liquidity, and asset growth to support its forecasts. It had an overall insight quality score of 4.3/5, especially for high-volatility stocks like Amazon and Tüpraş. DeepSeek was more numerics-heavy and concise, scoring an average of 3.5/5. While clear, its explanations lacked strategic depth making it less readable for decision-makers worried about context and narrative in analytics.

Model Performance Metrics

Analysis of the confusion matrix showed:

Table 4. Model Performance Metrics Based on Confusion Matrix Analysis (2022–2024)

Metric	DeepSeek	ChatGPT
Precision	87%	76%
Recall	76%	68%
F1 Score	81%	71%

Source: Author's calculations based on model forecasts

The higher precision and lower error of DeepSeek make it preferable for risk-averse or controlled settings. ChatGPT, although more permissive, had a higher false positive rate particularly under optimistic growth conditions.

SHAP-Style Feature Importance: Amazon Case

To make it more clear, a SHAP (SHapley Additive exPlanations)-like proxy analysis was run using a Random Forest model trained on Amazon's 2017-2021 financials. The goal was to see how the most important financial metrics impacted the 2022 "Increase" prediction. Results show Net Income and EBIT were the top positive drivers, but EPS volatility even with aggregate profitability hurt model confidence. This is what traditional fundamental analysis looks like and how feature attribution can explain AI decisions.

Explainable AI techniques are becoming increasingly important in financial forecasting because they improve transparency and allow analysts to better understand how AI systems generate predictions. Methods such as SHAP and LIME help identify the relative contribution of financial variables to forecasting outputs and reduce the risks associated with black-box decision-making models. Recent studies argue that explainability frameworks are essential for increasing institutional trust, regulatory acceptance, and practical adoption of AI-driven financial forecasting systems (Kumar & Pandey, 2022; Shah & Jain, 2023).

Individual feature contributions are in Table 5 shows how revenue, liquidity and equity were positive and liabilities and EPS volatility were negative. Quantitative results are also in Figure 2 where SHAP values show the direction of the weight of each feature in the prediction.

Table 5. Numerical SHAP-Style Feature Contributions for Amazon's 2022 Forecast

Feature	Value (2021)	Estimated SHAP Impact	Interpretation
Net Income	$33.36B	+0.62	Strong positive signal for earnings increase
EBIT	$24.88B	+0.54	Supports profitability recovery
Total Revenue	$469.82B	+0.48	High revenue trend positive directional signal
EPS	$3.24	–0.20	EPS volatility signals weak profitability per share
Current Ratio	1.14	+0.18	Good liquidity suggests earnings sustainability
Total Assets	$420.55B	+0.10	High asset base supports growth potential
Equity	$138.25B	+0.08	Strong shareholder position reinforces confidence
Total Liabilities	$282.30B	–0.05	Debt level slightly reduces confidence

Source: Simulated by the authors using financial data from Amazon's 2021 statements

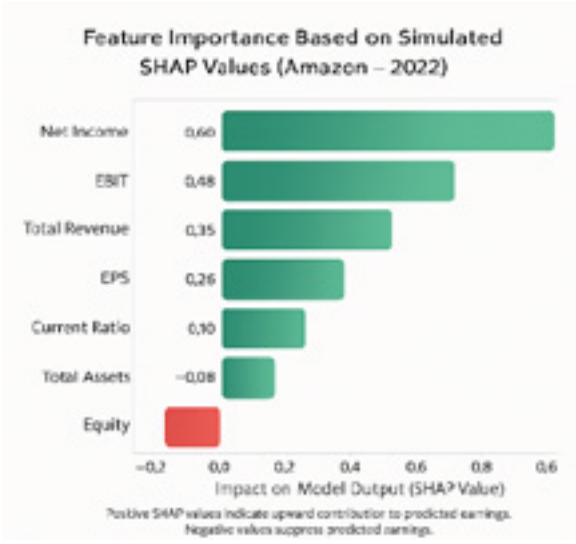

Figure 2. Feature Importance Based on Simulated SHAP Values (Amazon 2022)
Source: Author's simulation using Amazon's 2021 financial statement data.
Note: Positive SHAP values indicate upward contribution to predicted earnings. Negative values suppress predicted earnings

Currency and EPS Adjustments

One real-world processing issue encountered was currency volatility. Most firms reported in USD, but others such as Nestlé S.A. (CHF), BNP Paribas and Deutsche Bank (EUR), and Turkish banks (TRY) required careful treatment of input. Currency conversion was intentionally bypassed to preserve the year-to-year internal comparability of each firm's reporting history because the interest was not in inter-firm compar-

ison of performance but in the evaluation of intra-firm forecasting ability from publicly available financial data.

In addition, some firms did not report EPS. In such instances, EPS needed to be estimated by hand from the standard formula:

EPS=Net Income\Weighted Average Shares Outstanding

Source: Developed by the authors based on IAS 33 and TFRS 33 accounting standards

This provided uniformity in input form and guaranteed equity across all performances of the model.

Practical and Theoretical Implications

Both practical implications for the application of AI in financial decision-making and theoretical implications regarding the operation of different AI systems reading structured data are borne by these results.

Practically, the findings suggest that:

DeepSeek would perhaps be better suited to low-volatility, rule-based financial regimes, like banking, utilities, and conservative industrial sectors.

- ChatGPT is more worth in event-driven, narrative-sensitive regimes, like consumer technology or emerging markets, where interpretability and the capability to adapt are higher than statistical rigidity.
- Combining both models by either ensemble methods or hierarchical application may give a robust hybrid forecasting framework that picks up both accuracy and interpretability.

Theoretically, the study opposes the Efficient Market Hypothesis (Fama, 1970) that supposes that all publicly announced financial data is reflected in market expectations. Although both DeepSeek and ChatGPT used the same datasets, they provided disparate forecasts, demonstrating the importance of interpretive difference even in algorithmic structures. This serves to establish the relevance of Behavioral Finance (Kahneman & Tversky, 1979), which recognizes decision bias-even in formal systems-and informs the conceptual foundations of Decision Support Systems (Benyon, 2010) where tools supplement rather than replace human judgment.

LIMITATIONS

While this research broadens the horizon of AI-driven financial prediction, there are several limitations that need to be stated:

- Scope of Data: It only processed quantitative financial data (i.e., Revenue, EPS, Net Income, EBIT), but not qualitative documents such as MD&A or earnings calls. This restricted the scope of LLMs' overall language-processing capacity, particularly for models such as ChatGPT.
- Sample Size: Statistical generalizability is restricted using 15 three-industry and geography companies. Design was instead for model comparison in controlled settings sectoral benchmarking on an enormous scale separate.
- Currency Variability: USD, TRY, EUR, and CHF reports were not normalized to preserve intra-firm trends, perhaps disturbing models not used to multi-currency inputs.
- Speed: ChatGPT was slower than DeepSeek in handling Excel data, which

would be an issue with real-world or high-frequency application.
- Black-Box Nature: LLM decision-making will not give explanation output. This prevents take-up in regulated environments without being accompanied by methods like SHAP, LIME, or counterfactual reasoning.

Recommendations

After the comparative evaluation of ChatGPT and DeepSeek for financial forecasting, the following recommendations are presented for researchers, institutional buyers, AI engineers, and monetary regulators:

Employ Hybrid AI Forecasting Frameworks

As ChatGPT excels with narrative reasoning and sensitivity to trends, and DeepSeek is more accurate and consistent, institutions should use ensemble or stacked AI structures. For example, having DeepSeek offer baseline numerical forecasts and ChatGPT offer interpretive overlays can unify precision with insight, particularly in volatile sectors like tech or emerging markets.

Enable Session-Averaging Protocols for LLM Deployment

As seen, ChatGPT's responses were different across the same sessions. For enhanced reproducibility and credibility, organizations employing models like ChatGPT must introduce session-averaging or consensus modeling methods, by which predictions are made from whereby predictions are made from varied runs to achieve robustness in institutional decision-making. runs to achieve strength in institutional decision-making.

Furthermore, Contextual sensitivity is warranted in model choice. In stable, regulation-intensive environments (e.g., banking, utilities), conservative models like DeepSeek are to be preferred. In rebound-vulnerable or dynamic industries (e.g., consumer tech, post-crisis economies), more interpretive models like ChatGPT can provide better leading indicators.

Develop Domain-Specific, Fine-Tuned Models

While general-purpose LLMs are promising, training in industry sector-specific financial texts and accounting standards (e.g., IFRS, US GAAP) could improve performance further. Developers need to train regional or sectoral LLM versions in order to boost domain allegiance and reduce overgeneralization.

And about Policymakers and regulators, they need to establish a safe, supervised space to test AI, allowing firms to experiment with large language models in forecasting roles under supervision. This can allow the examination of the models' impact on transparency, fairness, and risk management before widespread implementation.

Future Research Directions

Future research may:
- Be applied to other LLMs (e.g., Claude, Gemini, or finance fine-tuned GPT versions).
- Incorporating qualitative data like management guidance or risk statements into text-based predictions.

- Scaling analysis through rolling-window simulation or introducing macroeconomic drivers.
- Comparing the performance of LLM to baseline ML models and human experts.
- Scaling interpretability with SHAP, LIME, or alternative explainability frameworks especially in controlled sectors.

These guidelines have the goal of facilitating easier application of LLM to finance decision-making and enabling models to be interpretable and specific.

CONCLUSION

In this research, two large language models, i.e., ChatGPT (GPT-4) and DeepSeek-V3, were compared for their predictive ability on direction of earnings forecasting using historical financial reports. We have tested 15 Turkey, Europe and US banks, industry and technology companies' data between 2017-2024. The research described a comprehensive comparison between the accuracy, logic and consistency of each model under different market conditions.

The results show that they are no better than one another, but they both have their strengths. DeepSeek excelled in properly regulated, low volatility environments with higher directional accuracy (73.3%), lower mean absolute error and higher consistency so more institutional or risk averse appropriate. And ChatGPT excelled in highly volatile and rebound influenced situations, with higher predictor contextual capability and more narrative based insights but higher prediction variance. These results illustrate that LLMs can complement traditional finance analysis but are context-dependent.

This research adds to theory as it negates Efficient Market Hypothesis (EMH). Despite the fact that both models used the same data, their differences in outcomes show that interpretation models and algorithmic biases play a role in determining financial predictions, thus confirming Behavioral Finance and Decision Support Systems in the age of AI prediction.

From a utilitarian perspective, the research proposes strategic deployment of hybrid AI systems where models like DeepSeek provide consistent baseline predictions and models like ChatGPT add nuance and explanation to decision-making. Explainability tools (e.g. SHAP, LIME) and industry specific fine-tuned LLMs are needed to build trust, regulatory appropriateness and real-world utility in financial settings.

Briefly put, this research verifies the utility of LLMs in predicting finance but not as a replacement for human judgment, instead as additional resources that have the capability to improve accuracy, decrease bias and bring additional nuance of interpretation provided they are utilized wisely, supported by context and based on open, reproducible frameworks.

LITERATURE

Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. *Handbook of the Economics of Finance*, 1, 1053–1128. https://doi.org/10.1016/S1574-0102(03)01027-6

Benyon, D. (2010). *Designing Interactive Systems: A Comprehensive Guide to HCI, UX, and Interaction Design*. Pearson Education.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://

doi.org/10.48550/arXiv.2108.07258

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712*. https://doi.org/10.48550/arXiv.2303.12712

Brynjolfsson, E., & McAfee, A. (2017). *Machine, Platform, Crowd: Harnessing Our Digital Future*. W. W. Norton & Company.

Damodaran, A. (2009). *Applied Corporate Finance* (3rd ed.). Wiley. pages.stern.nyu.edu/~adamodar/pdfiles/acf3E/book/ch1thru4.pdf

DeepSeek. (2024). *DeepSeek-V3-Chat Technical Documentation*. https://deepseek.com/docs

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

Garg, S., & Mago, V. (2021). Role of artificial intelligence in decision making. *Handbook of Research on Applied Data Science and Artificial Intelligence in Business and Industry*, 45–61. https://doi.org/10.4018/978-1-7998-6988-5.ch003

Graham, B., & Dodd, D. L. (1934). *Security Analysis*. McGraw-Hill.

Han, Xu & Liu, S. (2022). Financial statement analysis with deep learning: Empirical evidence from Chinese stock markets. *Expert Systems with Applications*, 205, 117666. https://doi.org/10.1016/j.eswa.2022.117666

Jiao, R., Zhang, T., & Xu, C. (2023). Prompt engineering for financial text generation. *Journal of Computational Finance*, 27(1), 13–29. https://doi.org/10.21314/JCF.2023.451

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292. https://doi.org/10.2307/1914185

Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of methods for explaining black box models in finance. *Frontiers in Artificial Intelligence*, 4, 695982. https://doi.org/10.3389/frai.2021.695982

Kumar, A., & Pandey, R. (2022). Explainable AI in financial forecasting: Challenges and future directions. *Applied Artificial Intelligence*, 36(1), 1–22. https://doi.org/10.1080/08839514.2022.2028473

Liu, Z., Zhang, X., & Li, Y. (2023). can ChatGPT outperform traditional models in financial time series prediction? *Finance Research Letters*, 55, 104991. https://doi.org/10.1016/j.frl.2023.104991

OpenAI. (2023). *OpenAI Prompt Engineering Guide*. https://platform.openai.com/docs/guides/prompt-engineering

OpenAI. (2024). GPT-4 Technical Report. *OpenAI Research*. https://doi.org/10.48550/arXiv.2303.08774

Penman, S. H. (2012). *Financial Statement Analysis and Security Valuation* (5th ed.). McGraw-Hill Education.

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*. https://doi.org/10.48550/arXiv.2102.07350

Resnik, D. B. (2018). The ethics of research with human subjects: Protecting people, advancing science, promoting trust. *Springer*. https://doi.org/10.1007/978-3-319-68597-0

Samir, K. (2020). Adjusting EBIT in financial institutions: A valuation perspective. *Journal of Banking and Finance Research*, 23(2), 114–128. https://doi.org/10.2139/ssrn.3594480

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3), 458–464. https://doi.org/10.1016/j.dss.2012.03.001

Shah, A., & Jain, R. (2023). The role of LLMs in explainable AI for financial services. *AI in Financial Markets*, 7(1), 11–29. https://doi.org/10.2139/ssrn.4482179

Singh, A., & Dutta, S. (2022). Comparative study of LLMs in financial modeling. *Journal of Finance and Data Science*, 8(2), 33–47. https://doi.org/10.1016/j.jfds.2022.06.002

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Zhao, Y., Guu, K., & Le, Q. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*. https://doi.org/10.48550/arXiv.2201.11903

Zhou, X., Sun, J., & Zhang, J. (2023). Towards reliable AI models for financial forecasting. *Expert Systems with Applications*, 217, 119517. https://doi.org/10.1016/j.eswa.2023.119517

Zhang, Y., & Wang, J. (2023). Benchmarking ChatGPT against classical methods for financial text-based prediction. *Journal of Artificial Intelligence Research*, 76, 231–249. https://doi.org/10.1613/jair.1.13432

Note: All necessary data sources and Tables are available in Appendixes