https://doi.org/10.31891/csit-2026-2-22

UDC 004.89: 336.7

ARTIFICIAL INTELLIGENCE IN FINANCIAL TECHNOLOGY: METHODS, APPLICATIONS, AND CURRENT DEVELOPMENTS

Serhii SAVCHENKO
PhD, Doctorate of Computer Engineering & Information Systems Department,
Khmelnytskyi National University
Lecturer of Faculty of Computer Science, Physics and Mathematics,
Kherson State University
https://orcid.org/0000-0001-9706-5334
e-mail: savchenko.serhii@gmail.com

This study systematises modern methods of generative artificial intelligence, specifically large language models (LLMs), and analyses approaches for their application in the financial technology sector. It provides a summary of existing strategies for using LLMs in FinTech, including zero-shot, few-shot, fine-tuning, Retrieval-Augmented Generation (RAG), and training models from scratch. A comparative analysis of their cost and implementation complexity was performed, identifying the most suitable LLM integration options depending on the application task. The paper presents an algorithm for automatic investment portfolio rebalancing, which combines classical Markowitz Portfolio Theory (MPT), price forecasting using LSTM networks, and technical analysis signals. An extended version of the rebalancing algorithm is proposed, supplementing traditional quantitative methods with two LLM components: a market sentiment analysis module and a financial statement processing module. Integrating these components enables the processing of unstructured data, such as financial news, social media posts, and quarterly or annual corporate reports. Using such data significantly expands the input datasets for price forecasting models, which can improve the quality of investment decisions. Based on the analysed scientific publications, it is shown that combining technical and fundamental financial indicators with market sentiment assessment helps to increase the accuracy of price forecasting for financial instruments. The paper demonstrates the potential for using the proposed investment portfolio rebalancing method in automated financial advisory systems (Robo-Advisors). The main limitations of the study are highlighted, in particular the need to test the rebalancing algorithm in practice using real market data. Directions for further research are identified, relating to the experimental testing of the proposed model on historical data from various periods and the subsequent optimisation of LLM components based on the results of the experiments.

Keywords: generative artificial intelligence; large language models; machine learning; financial technologies; investment portfolio rebalancing; market sentiment analysis.

Received: 04/04/2026
Accepted: 11/05/2026
Published: 31/05/2026

Introduction

Due to the active development of artificial intelligence (AI) technologies, the financial sector, like many others, is currently undergoing a significant transformation. While the application of AI in financial technology (FinTech) has developed gradually over recent decades, a major leap occurred in the 2020s, primarily driven by the emergence of large language models (LLMs). An analysis of recent scientific publications demonstrates a growing academic interest in the use of AI within finance. In recent years, there has been a clear trend towards an increasing number of studies focused on the practical implementation of AI technologies in FinTech products [1, 2].

The evolution of AI usage in finance can be divided into several key stages [1]. Between 1990 and 2015, classical statistical models and machine learning methods were predominantly used, such as support vector machines (SVM), regression models, and decision trees. These were applied to tasks like price forecasting, credit risk assessment, and fraud detection. From 2015 to 2019, deep learning (DL) was actively utilised. Long Short-Term Memory (LSTM) recurrent networks, convolutional neural networks (CNN), and ensemble methods proved effective for time-series forecasting and automated trading [3]. However, these approaches were mainly focused on structured data, such as prices, trading volumes, transaction counts, and formalised financial statements.

The current stage is marked by the active implementation of generative AI (GenAI), primarily Large Language Models (LLMs). Unlike deep learning, LLMs allow for working with unstructured data (financial reports, news, social media posts, thematic forums, regulatory documents, etc.). As noted in study [4], LLMs, specifically the GPT series and FinBERT, are already showing good results in text sentiment analysis and document summarisation. Research [5] emphasises that the ability of LLMs to synthesise diverse information from news, corporate communications, and financial reports significantly accelerates investment decision-making. Work [6] demonstrates that GPT-4 and Llama 3 models, using correctly selected prompts, reduce the time for credit analytics of small and medium enterprises from weeks to minutes.

Due to the continuous growth of unstructured financial data volumes, traditional quantitative methods are unable to process them effectively [7]. Competitive pressure on financial companies regarding the speed and quality of analytics is increasing rapidly. FinTech startups are integrating AI tools into their software architecture during the early stages of development [1]. Therefore, there is a critical need to combine traditional quantitative methods with qualitative text analysis to enhance the input data for decision-making models [8].

The study [9] presents an algorithm for rebalancing an investment portfolio, which combines Markowitz portfolio theory for initial portfolio construction with technical analysis signals (MACD, RSI, S&R) for making trading decisions. The concept proposed in this paper involves supplementing this approach with generative AI tools to expand input data with qualitative signals from unstructured information sources.

The purpose of this study is to systematise and analyse modern generative AI methods for the FinTech industry, considering implementation costs and evaluation metrics. Furthermore, it aims to expand the investment portfolio rebalancing algorithm by integrating LLM components for the unstructured data analysis.

Literature review

The application of AI in the financial sector began with the use of traditional machine learning (ML) methods. Among these, the most popular were the SVM method, decision trees, and ensemble approaches such as Random Forest and Gradient Boosting (XGBoost). These algorithms were successfully applied to address three main classes of financial tasks: asset price forecasting, credit risk assessment, and fraud detection.

The systematic review [3] states that SVM systems have been widely used to predict the direction of financial asset prices, showing a classification accuracy of 55-65% even for cryptocurrency assets. The study [10] demonstrates that Random Forest and XGBoost significantly improved the results of traditional models in asset valuation and credit scoring tasks. In particular, ensemble methods show resistance to overfitting and effectively process unbalanced data sets, which is a typical problem in fraud detection. However, with the increasing volume of unstructured financial data (news, analytical reports, social media data), classic ML methods face certain limitations. They are generally unable to process sequential data with complex temporal dependencies and adapt poorly to the non-stationarity of financial markets. For example, SVM approaches that rely on the assumption of repeating market conditions prove to be ineffective in "black swan" scenarios [11].

During the 2010s, the development of graphics processing unit (GPU) computing technology and the accumulation of vast datasets led to the active implementation of deep learning in many applied fields, particularly in the financial sector. RNNs and their advanced variants, such as LSTM and GRU (Gated Recurrent Unit) networks, significantly expanded the possibilities for time series modelling in forecasting tasks. In study [12], a wide range of DL models for financial asset management was systematised. LSTM networks demonstrated high efficiency in predicting stock volatility and returns due to the gate mechanism, which selectively preserves or discards information over long sequences. Hybrid CNN-LSTM architectures combined the advantages of convolutional networks for detecting local patterns with recurrent layers for modelling temporal dependencies [11]. In study [9], an LSTM network is used to predict asset closing prices over a one-month horizon. The investment portfolio formed based on these forecasts, constructed according to Markowitz theory, demonstrates better results compared to the "buy-and-hold" strategy even during a market recession. This approach illustrates a productive combination of deep learning methods with classical portfolio theory.

The study [13] systematised the fields of application for CNN, RNN, and hybrid models regarding price movement forecasting in financial markets. Research [6] presents data demonstrating the advantages of hybrid LSTM+ARIMA approaches over single models in volatility forecasting tasks. Despite progress in applying DL, recurrent architectures possess certain structural limitations. Such neural networks process sequences incrementally, which results in high computational costs for long series, and they face difficulties in maintaining context over very large intervals.

A significant milestone in the development of AI was the emergence of the Transformer architecture in 2017, which marked the beginning of a new paradigm in natural language processing (NLP). Transformers use a self-attention mechanism to model the relationships between words in parallel, which has significantly accelerated training on large datasets [14]. Thus, transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have opened up new possibilities for financial data analysis.

Among the key financial-oriented LLMs, BloombergGPT is particularly noteworthy. It is the first model trained from scratch on 363 billion financial tokens combined with 345 billion public tokens. Training required 1.3

million GPU hours on A100-class hardware [8]. The FinGPT model represents a fundamentally different strategy [14]. FinGPT is an open financial LLM that uses a fine-tuning method based on low-rank adaptation (LoRA). This allows the model to be adapted for financial tasks for less than 300 US dollars, reducing adaptation costs approximately 1000 times compared to full fine-tuning. FinBERT is a specialised version of the BERT LLM model, further trained on an array of financial news and analytics. It established a new standard for the analysis of financial texts [4].

The study [8] proposes a four-level framework for selecting an LLM implementation strategy based on available resources: Level 1 – zero-shot application, where costs are limited to API request fees; Level 2 – few-shot learning using a limited number of examples, with minimal additional costs; Level 3 – fine-tuning using LoRA or instruction tuning (costs depend on the amount of data); Level 4 – training from scratch, which requires millions of GPU hours and substantial financial resources.

Based on the researched sources, four main functional areas for applying LLMs in finance can be identified:

- Algorithmic trading and portfolio optimisation. LLMs are capable of processing unstructured textual information (market news, financial reports, social media) to generate trading signals [5]. LLM-based systems go beyond traditional quantitative analysis by understanding context, detecting sarcastic remarks, and interpreting complex financial terminology. In portfolio optimisation tasks, these models provide a more accurate analysis of potential risks compared to traditional approaches [5].
- Study [15] describes the use of LLMs for predicting mergers and acquisitions (M&A), bankruptcies, and market movements. Specifically, the RiskLabs framework integrates textual data from management earnings calls, market time series, and news for comprehensive financial risk forecasting.
- LLMs are actively used in fraud detection and ensuring regulatory compliance. GPT-oriented models demonstrate effectiveness in identifying fraudulent patterns in payment systems through the analysis of temporal and contextual sequences [15].
- Robo-Advisors (RA) and LLM-based chatbots are transforming the interaction between financial institutions and clients. Research [16] shows that LLMs allow RA services to provide personalised investment recommendations that consider risk tolerance and market dynamics. Systems such as GPTQuant use few-shot learning to generate and analyse investment strategies [15].

An analysis of the existing literature allows for the identification of the following areas that remain insufficiently researched:

- There is a lack of systematic research into hybrid LLM+LSTM architectures for financial forecasting. Although studies [11, 12] demonstrate the benefits of hybrid approaches within classical DL methods, and works [4, 16] show the potential of LLMs for financial tasks, the combination of transformer and recurrent neural network architectures in a single framework for time series forecasting remains poorly explored.
- The comparative cost assessment of different LLM implementation strategies in financial organisations is fragmented. While study [8] proposes a conceptual four-level framework and [14] provides specific values for FinGPT, there are no empirical comparative studies that consider the total cost of ownership (TCO) given the specifics of financial regulation.
- The methodology for quantifying the effectiveness of generative AI in financial applications remains unstandardised. Existing benchmarks described in [4, 17] focus mostly on classification tasks, whereas uniform evaluation metrics for generative financial tasks (report generation, trading strategy generation, investment decision explanations) are missing.

Methods for implementing generative artificial intelligence in FinTech

The increasing diversity of approaches to applying LLMs within the financial sector requires a systematisation of existing solutions. Let us consider the four-level taxonomy of LLM approaches proposed in the study [8].

Level 1: Zero-shot and Few-shot learning. The most accessible option is using ready-made models (GPT, Claude, Gemini) without additional training, according to the zero-shot or few-shot principle. This approach does not require costs for developing a training set, and the deployment cost is limited to API call charges. These range from $0.002 to $0.12 per 1,000 tokens for third-party services (OpenAI, Anthropic, Google), or from $0.006 to $0.037 when deploying open models independently [8]. Primary applications include financial report analysis, compliance checks, and generating responses to client queries. The advantages are minimal implementation costs and rapid deployment; however, the accuracy of such solutions may be insufficient for complex, domain-specific tasks.

Level 2: Fine-tuning of open models. If a zero-shot approach does not provide satisfactory quality, the next step involves fine-tuning open models (such as LLaMA or Gemma) on financial data. A notable example is the FinGPT model [14], which utilizes the Low-Rank Adaptation (LoRA) technique. To fine-tune approximately 50,000 samples, the cost of one training cycle is less than $300. The FinMA-7B model (part of the PIXIU project) is more resource-intensive, requiring 8 A100 GPUs (40 GB each) to train on 136,000 instruction samples. The FinMA-30B model already requires 128 such GPUs [8]. Overall, fine-tuning costs vary from $4 to $360,000 depending on the model size and data volume. The primary tasks for such models include market sentiment classification, Named Entity Recognition (NER), and financial text analysis.

Level 3: Tool-Augmented Generation and RAG. Provided that fine-tuning does not solve the task, it is appropriate to supplement the LLM with external tools. The RAG (Retrieval-Augmented Generation) approach involves obtaining relevant documents from external repositories and including them in the query context. Deploying such systems involves accompanying costs for the development of the tools themselves [8]. Typical applications include financial chatbots, customer support systems, and Robo-Advisory systems with access to current market data.

Level 4: Training from Scratch. This is the most resource-intensive level, involving the training of an entirely new LLM on a massive domain-specific dataset. A primary example is BloombergGPT—a model with 50 billion parameters based on the BLOOM architecture, trained on 363 billion financial tokens combined with 345 billion tokens from general sources. The total costs amounted to 1.3 million A100-hours, which at AWS rates ($2.3/hour) equals approximately $3 million. Overall, expenses for this level exceed $5 million, and the volume of training data reaches hundreds of billions of tokens. The target area of application consists of specialised financial analytical systems at a corporate scale.

Table 1

Comparison of LLM application levels in finance (based on [8])

Level	Approach	Development Cost	Data Volume	Inference Cost ($/1K tokens)
1	Zero-shot / Few-shot (third-party API)	$0	0	$0.002–$0.12
1	Zero-shot / Few-shot (open model)	$0	0	$0.006–$0.037
2	Fine-tuning (open model)	$4–$360,000	10,000 –12,000,000	$0.002–$0.12
3	Tool-Augmented Generation / RAG	Cost of tools	–	$0.002–$0.12
4	Training from scratch	>$5,000,000	>700,000,000	$0.002–$0.12

The efficiency of LLMs in a financial context is driven by the self-attention mechanism, which enables the model to process interdependencies between arbitrary positions in the input sequence in parallel, regardless of the distance between them [5]. This is particularly important in financial settings, as the model can identify causal links between economic events (such as report publications or central bank decisions) and asset price movements described in different parts of a long text. Thanks to the transformer architecture, LLMs overcome the limitations of previous recurrent networks (LSTM, GRU) when dealing with long-range dependencies [4].

The Chain-of-Thought (CoT) technique involves formulating a request in such a way that the model demonstrates step-by-step logic before providing an answer. Research [5] shows that CoT prompting significantly improves the quality of performing financial tasks compared to simple zero-shot prompting. Specifically, the application of CoT allows GPT-4 to outperform traditional models during the analysis of financial statements and when forecasting company earnings, even without specialised fine-tuning [16]. This technique is also successfully used in multi-agent systems to break down complex financial tasks into subtasks and for step-by-step decision-making [16].

Low-Rank Adaptation (LoRA) is an effective method of parameter-efficient fine-tuning, where only low-rank adapter matrices are updated instead of all model weights. In the FinGPT system, the application of LoRA allows for the reduction of training parameters from 6.17 billion to 3.67 million, which is approximately a 1000-fold decrease [14]. This significantly lowers GPU memory requirements (by about 50%) and training time, making fine-tuning accessible even without large computing clusters. The FinGPT model trained using LoRA achieves a financial sentiment classification accuracy of 82.1%, compared to 63.4% for ChatGPT in zero-shot mode [14].

One of the primary limitations of LLMs is hallucination – the tendency to generate plausible but factually incorrect responses. In the financial sector, where accuracy is critically important, Retrieval-Augmented Generation (RAG) is becoming the standard mechanism for addressing this issue. RAG integrates LLMs with vector databases containing tokenised domain information. Before generating a response, the system retrieves relevant documents (reports, regulations, news) and passes them along with the user's query in a single context [14]. The application of RAG increases the reliability of responses and allows LLMs to operate with up-to-date data that extends beyond the model's training window. Systems such as MarketSenseAI 2.0 already combine RAG with LLM agents to analyse SEC filings, earnings reports, and institutional records [16].

Thus, the range of technical approaches for applying generative AI in FinTech covers solutions from the simple use of ready-made APIs to the full training of specialised corporate-level systems. The choice of the optimal approach is determined by the balance between accuracy requirements, available computing resources, and the nature of the financial tasks being addressed.

Applications of generative artificial intelligence in FinTech

Sentiment analysis of financial texts is one of the most established areas of NLP application within FinTech. The evolution of NLP methods has progressed from dictionary-based approaches (specifically models that calculate the frequency of positive and negative words) through statistical classifiers (such as Naive Bayes and SVM) to neural network architectures based on transformers [5].

Currently, the dominant tool for classifying the sentiment of financial texts into categories (positive / neutral / negative) is the FinBERT model. It is trained on a specialised financial corpus and can account for specific financial terminology. Alongside this, zero-shot learning approaches based on general LLMs are actively used. On the FiQA-SA (Financial Question Answering – Sentiment Analysis) dataset, the GPT-4 model in zero-shot mode achieves a classification accuracy of around 79% [16]. Compared to lexical and statistical methods, LLM-based approaches demonstrate significantly higher accuracy, especially when processing complex linguistic structures such as irony and double negatives.

Predicting the prices of financial assets is one of the most extensively researched applications of AI in fintech. Traditional approaches, based solely on technical indicators – RSI (Relative Strength Index), MACD (Moving Average Convergence Divergence) and others – are gradually being replaced by hybrid architectures that combine LSTM with language models [11].

The study [18] describes an approach where an LLM model analyses unstructured financial data to generate investor sentiment features, which are then input into an LSTM to predict stock dynamics. Results from such hybrid LSTM + sentiment analysis systems show that trend prediction accuracy reaches 92%. Specific LLM architectures for predicting market trends, which combine quantitative factors (moving averages, option volumes) with text signals from real-time news, demonstrated an accuracy of up to 95% for certain assets [18].

Another promising direction is the application of reinforcement learning (RL). Frameworks such as FinRL allow for the training of agents for automated trading. In this process, the agent receives environmental signals, such as market data and portfolio position, and learns to maximise long-term rewards, like the Sharpe ratio, without explicit programming of trading rules. An increasing number of studies integrate multi-agent LLM frameworks to simulate investor behaviour and adapt to market conditions in real time. This confirms a steady trend towards the convergence of LLMs with RL in modern algorithmic trading systems [16].

Markowitz's classic Modern Portfolio Theory (MPT), despite its mathematical precision, has significant limitations. It relies entirely on retrospective numerical data, such as expected returns and covariance matrices. It ignores qualitative factors like geopolitical events or changes in market sentiment and assumes static asset weights. These limitations become critical in unstable market conditions. LLMs address these weaknesses by analysing unstructured data, including market reports, news, and financial statements. Combining quantitative signals with qualitative insights extracted by LLMs allows for more adaptive assessments of asset risk and return [5]. Study [19] demonstrated that, in terms of recommendation quality for three different investor profiles, ChatGPT outperformed 14 out of 17 tested classic Robo-Advisors.

Modern Robo-Advisor systems with an LLM component go beyond automated asset allocation. They are able to personalise client interaction using natural language, dynamically explain algorithmic decisions, and adapt to changes in market conditions [20]. An important direction is also ESG scoring (Environmental, Social, Governance) based on LLMs. Contemporary models can automatically evaluate companies according to these parameters, analysing public documents and news feeds, which previously required significant analytical resources [21].

Automated analysis of financial statements represents another area where LLMs demonstrate transformative potential. Documents such as 10-K (annual reports) and 10-Q (quarterly reports) submitted to the US Securities and Exchange Commission (SEC) typically exceed 100 pages. These documents contain complex disclosures that require significant analytical resources [16]. LLMs are capable of automatically extracting key metrics, identifying early signs of financial distress, and analysing changes in the sentiment of corporate communications over time [5].

Thus, the areas of application for generative AI discussed here indicate that LLMs are evolving from a supporting tool into a key component of the financial technology infrastructure, operating alongside or in place of traditional analytical systems and human analysts.

Application of LLMs for investment portfolio rebalancing

In the study [9], the authors developed an investment portfolio management algorithm. The initial investment portfolio is constructed based on Markowitz's portfolio theory (MPT). Historical data (monthly closing prices) and forecast prices for the following month (using an LSTM model) are used as input for the MPT. The rebalancing algorithm checks the signals of technical indicators (MACD, RSI, S&R) on a weekly basis, compares the current weights with those in the newly constructed optimal portfolio, and, if there is a discrepancy, executes buy-sell transactions. Backtesting using data from various time periods confirmed that the rebalancing strategy based on S&R and RSI indicator signals delivers better results compared to the passive 'buy and hold' strategy [9]. The flowchart of the rebalancing algorithm is shown in Fig. 1.

The presented rebalancing mechanism has a fundamental limitation. The LSTM model used for forecasting financial instrument prices relies solely on historical share price data, ignoring fundamental company information and market sentiment signals. As shown in study [12], many modern portfolio management systems only use historical and technical data sources, while combining fundamental and technical analysis demonstrates better results. An optimal investment portfolio rebalancing model should, among other things, consider fundamental changes in company activities (such as changes in revenue, debt pressure, or management changes) and mass shifts in market sentiment.

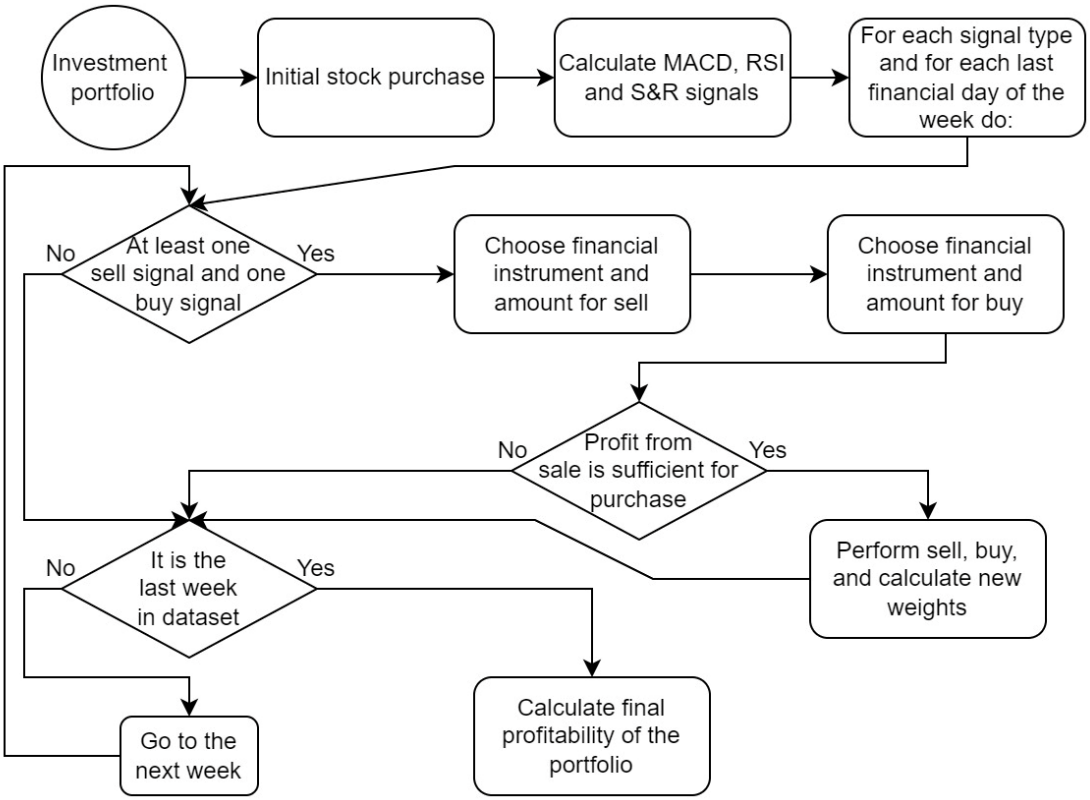

Fig. 1. Algorithm for the automated rebalancing of an investment portfolio

To address this limitation, an extended hybrid architecture is proposed, which supplements the base algorithm [9] with two LLM components (Fig. 2).

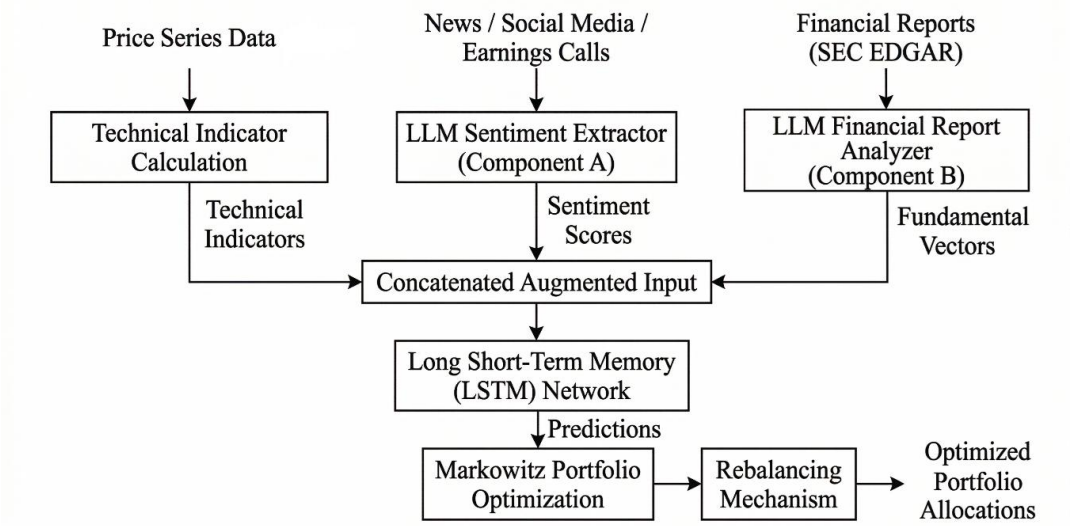

Fig. 2. Schematic representation of the proposed hybrid architecture

LLM-based Sentiment Feature Extractor (Component A). Input data consists of news streams from financial media (Reuters, CNBC, Yahoo Finance), social media posts (X, Reddit), and transcripts of quarterly earnings calls. The use of a fine-tuned FinGPT or FinBERT model is proposed as the primary tool. According to research results [14], FinGPT, based on Llama-3.1-8B-Instruct using the LoRA method, achieves an accuracy of 82.1% and Macro-F1 = 80.9 on the financial sentiment classification task, significantly outperforming the base FinBERT (71.2% / 69.9). The key advantage of the FinGPT approach lies in the use of a market-oriented RLSP (Reinforcement Learning on Stock Prices) mechanism. This allows sentiment to be correlated with actual market consequences of news, rather than just subjective human labelling [14]. The output data consists of daily numerical sentiment scores in a three-dimensional class space (positive, neutral, negative) for each portfolio asset, forming a time series of sentimental indicators.

LLM-based Financial Report Analyser (Component B). The input data consists of quarterly (10-Q) and annual (10-K) reports of public companies according to US SEC (Securities and Exchange Commission) standards and IFRS (International Financial Reporting Standards) reporting, available via the public SEC EDGAR API. The methodological basis is a RAG pipeline based on open-source models from the LLaMA family with LoRA fine-tuning. The feasibility of the RAG approach is confirmed in study [14], which demonstrates that including external documents in the model's context increases the accuracy and relevance of answers. Furthermore, research results [16] show that the MarketSenseAI 2.0 system, built on a combination of RAG and LLM agents for processing SEC materials and earnings calls, achieved a cumulative return of 125.9% on S&P 100 stocks compared to 73.5% for the index. The output data of Component B includes: numerical vectors of fundamental features (revenue growth rate, debt load ratio, profitability indicators); and a structured text summary to ensure the interpretability of decisions. As noted in study [15], LLMs are capable of effectively processing long financial reports if a strategy of dividing the document into structural sections (management's discussion, financial highlights, business overview) is applied, which allows for overcoming the limitations of the models' context window.

Advanced LSTM input. It is proposed to concatenate three types of features into a single input vector: (1) price time series with technical indicators (daily frequency); (2) sentiment scores from Component A (daily frequency); (3) fundamental vectors from Component B (quarterly frequency). The difference in frequencies between price and sentiment data on one hand, and fundamental data on the other, represents a key technical challenge. To resolve this, the application of the forward-fill imputation method is proposed, which involves propagating the last known value and is standard practice for quarterly financial indicators in daily forecasting tasks.

The choice of specific LLM models should be based on cost-effectiveness criteria, taking into account daily (rather than real-time) rebalancing. GPT-4 in zero-shot mode provides high-quality analysis of financial reports without the need for fine-tuning, but is associated with significant costs for API requests and the transfer of sensitive data to external servers. FinGPT, on the other hand, is an open-source model [14] that allows for self-hosted deployment, which is critical for ensuring the confidentiality of client portfolio data. Research [8] emphasises that the application of LLMs in the financial sector requires particular attention to issues of confidentiality and compliance with regulatory requirements.

As noted in study [18], the combination of fundamental and sentiment signals with price data provides better predictive quality for medium-term horizons compared to using technical data alone. Price series on their own can only capture the effects of market events, whereas LLM components allow for the early detection of fundamental shifts before they are fully reflected in quotes. Increased debt burdens, declining profitability, or changes in a company's operational strategy—identified by Component B from quarterly reports—enable the LSTM model to adjust the asset's return forecast before these trends materialise in the price. Research [15] confirms that LLMs supplied with fundamental financial data can outperform traditional stock rating models in the accuracy of future earnings forecasts.

Social media and news reports reflect the collective expectations of market participants before these expectations are reflected in prices. According to study [8], including sentiment analysis based on LLMs as an additional signal in predictive models significantly improves their ability to detect shifts in market trends early. Finally, the contribution to interpretability is important: textual summaries from Component B allow for an understanding of which fundamental factors influenced the change in portfolio weightings. This meets the increasing regulatory requirements regarding the explainability of algorithmic investment decisions [16].

Thus, the proposed hybrid architecture develops the approach presented in [9], maintaining the proven Markowitz optimisation mechanism and weekly rebalancing, but fundamentally expands the LSTM forecasting information base by integrating LLM components for sentiment and fundamental analysis.

Evaluation methods and implementation costs of AI solutions in the FinTech sector

The study [8] proposes a four-level decision-making framework regarding the deployment strategy of LLM solutions, which allows for a balance between cost and efficiency. A comparative analysis of the four strategies is presented in Table 2.

Table 2

Comparison of LLM deployment strategies in financial applications

Deployment Strategy	Development Cost	Training Data Volume	Infrastructure Requirements	Time to Market
Zero-shot / Few-shot (third-party API)	$0	–	None (API access)	Days
Zero-shot / Few-shot (open-source, self-hosted)	$0	–	NVIDIA V100 / A100 class GPU	Weeks
Fine-tuning (third-party API)	$30–$30,000	10,000–12,000,000 samples	None (API access)	Months
Fine-tuning (open-source)	$4–$360,000	10,000–12,000,000 samples	NVIDIA A100/A6000 class GPU	Months
Training from scratch	$5,000,000+	700,000,000,000+ tokens	A100 cluster; 80 GB+ RAM	Years

In summary, it is recommended to start with a zero-shot approach and only move on to more resource-intensive options if the quality of the results is insufficient. Training an LLM from scratch (such as BloombergGPT) is justified in exceptional cases where existing solutions do not meet the requirements of a specific financial domain.

The correct choice of evaluation metrics is critical for validating AI systems in the financial sector, as different types of tasks require different quality criteria. For fraud detection and sentiment analysis tasks, the standard set includes Accuracy, Precision, Recall, F1-score, and AUC-ROC. However, as shown in studies [22, 23], in tasks with significant class imbalance (and fraud detection is exactly such an example), the Accuracy metric can be misleading. In these cases, preference should be given to AUC-PR, which focuses exclusively on the detection quality of the positive (fraudulent) class. Additionally, economically-oriented metrics are applied: Cost of False Positives – the cost of false alarms (inconvenience for clients, manual verification costs) and Cost of False Negatives – financial losses from missed fraudulent transactions [22].

For regression tasks (predicting asset prices and volatility), the most frequently used metrics are RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error). Directional Accuracy (DA), which represents the proportion of correctly predicted price movement directions, serves as an important additional metric. To evaluate trading signals, the Sharpe ratio of the predictive signal is used [24].

A comprehensive quality assessment of LLMs in a financial context must cover not only accuracy metrics but also dimensions such as fairness, robustness, and bias, as accuracy and efficiency must be supported by the meaningfulness of the decisions made.

Discussion of the results

The generalisation of the conducted review allows for the systematisation of four main LLM approaches in the form of a summary comparative characteristic (Table 3). An analysis of these characteristics confirms that zero-shot and few-shot approaches are optimal for new financial tasks or situations with limited annotated data, as they provide an acceptable level of quality at minimum cost. In contrast, fine-tuning is only appropriate when a sufficiently large industry-specific dataset and a clearly defined functional task are available.

Table 3

Comparative characteristics of LLM approaches for financial tasks

Method	Typical tasks	Advantages	Limitations
Zero-shot	Compliance audit, report analysis, customer responses	Minimal time to market (days); no need for labelled data	Lower accuracy on highly specialised tasks; dependence on API provider
Few-shot	Sentiment classification, financial event analysis	Rapid prototyping; flexibility	Limited customisation; sensitivity to example phrasing
Fine-tuning (LoRA)	Sentiment classification, NER, financial Q&A	High accuracy on domain tasks (FinGPT: 82.1% Accuracy); possibility of self-hosted deployment	Need for 10,000+ labelled samples; long preparation cycle (months)
Training from scratch	Full-cycle corporate financial systems	Maximum specialisation (BloombergGPT); no dependence on base models	Extremely high cost and duration; justified only in exceptional cases

A comparison of the theoretical rationale for the proposed hybrid architecture with the empirical results of similar approaches in the literature demonstrates its practical potential. Study [18] shows that hybrid LSTM + sentiment analysis systems, where the LLM model derives investor sentiment features from news and social media posts, achieve a market trend prediction accuracy of 92%, which significantly exceeds the results of the baseline LSTM, which relies solely on price series. Certain multi-factor LLM architectures, combining quantitative indicators with real-time text signals, have demonstrated accuracy of up to 95% for individual assets [18]. A key distinction of the architecture proposed in this work is the inclusion of an LLM-based Financial Report Analyzer, which generates vectors of fundamental features based on quarterly and annual reports (10-K/10-Q). This approach provides an additional advantage over purely sentiment-oriented systems, particularly for medium-term forecasting horizons.

The proposed rebalancing concept is based on the results of similar approaches in academic literature, but it has not been verified using real financial data. Conducting such testing represents a direct path for future research. It is important to note that the conclusions of this work rely on scientific publications, where the relevance of data is limited by their publication date. The field of LLMs is developing at an unprecedented speed, and specific accuracy or cost metrics for models cited in the analysed sources may change as new models emerge.

Conclusions

The conducted study systematised modern methods of generative artificial intelligence, specifically large language models, within the context of their application in financial technology. LLMs represent a fundamentally new class of tools in FinTech, moving beyond traditional machine learning due to their ability to process unstructured financial data without prior manual labelling. A comparative analysis of deployment strategies demonstrated that zero-shot and RAG approaches provide acceptable efficiency at minimal cost and are the optimal initial choice for most tasks. Meanwhile, fine-tuning is justified mainly for highly specialised classification tasks when a sufficient industry-specific data corpus is available.

The proposed hybrid architecture for extending an automatic investment portfolio rebalancing algorithm is theoretically sound. It integrates two LLM components: a sentiment feature extractor and a financial report analyser. This system can be implemented using existing open-source models. However, the study has several limitations that suggest directions for future research. The proposed concept is theoretical and analytical; it has not been verified with real market data, which is the immediate task for the next stage. The most promising areas for further study include agent systems and small specialized financial models designed for local deployment.

ADDITIONAL INFORMATION

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

In preparing this work, the author used DeepL Translate and Kagi Translate for translation, grammar checking, and rephrasing. The author also used the Elicit and NotebookLM services for literature search and preliminary analysis of academic sources. After using these tools, the author carefully reviewed and edited the content and assumes full responsibility for the content of this publication.

1. Vuković D. B., Dekpo-Adza S., Matović S. AI integration in financial services: a systematic review of trends and regulatory challenges. Humanities and Social Sciences Communications. 2025. Vol. 12. Article 562. DOI: https://doi.org/10.1057/s41599-025-04850-8

2. Mukthar K. P. J., Chauhan N., Al-Absy M. S. M. et al. Research dynamics in AI and fintech: a bibliometric investigation using R. Discover Internet of Things. 2025. Vol. 5. Article 19. DOI: https://doi.org/10.1007/s43926-025-00111-x

3. Dakalbab F., Abu Talib M., Nasir Q., Saroufi T. Artificial intelligence techniques in financial trading: A systematic literature review. Journal of King Saud University – Computer and Information Sciences. 2024. Vol. 36, Issue 3. Article 102015. DOI: https://doi.org/10.1016/j.jksuci.2024.102015

4. Nie Y., Kong Y., Dong X., Mulvey J. M., Poor H. V., Wen Q., Zohren S. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. arXiv. 2024. DOI: https://doi.org/10.48550/arXiv.2406.11903

5. Zhao H., Liu Z., Wu Z., Li Y., Yang T., Shu P., Xu S., Dai H., Zhao L., Jiang H., Pan Y., Chen J., Zhou Y., Zhang Z., Sun R., Mai G., Liu N., Liu T. Revolutionizing Finance with LLMs: An Overview of Applications and Insights. arXiv. 2024. DOI: https://doi.org/10.48550/arXiv.2401.11641

6. Yang R., Wang Y., Luo Y., Yang Z., Zong Z., Wu D. O. Recent Advances in Artificial Intelligence for Management and Financial Technology. Transactions on Artificial Intelligence. 2025. Vol. 1, No. 1. P. 139–152. DOI: https://doi.org/10.53941/tai.2025.100009

7. Rizinski M., Trajanov D. AI Agents in Finance and Fintech: A Scientific Review of Agent-Based Systems, Applications, and Future Horizons. Computers, Materials & Continua. 2026. Vol. 86, No. 1. P. 1–34. DOI: https://doi.org/10.32604/cmc.2025.069678

8. Li Y., Wang S., Ding H., Chen H. Large Language Models in Finance: A Survey. Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF 2023). New York : Association for Computing Machinery, 2023. P. 374–382. DOI: https://doi.org/10.1145/3604237.3626869

9. Savchenko S., Kobets V. Increasing Investment Portfolio Profitability with Computer Analysis Trading Strategies. Information and Communication Technologies in Education, Research, and Industrial Applications. Cham : Springer Nature Switzerland, 2023. P. 252–264. DOI: https://doi.org/10.1007/978-3-031-48325-7_19

10. El Alami S. E. A., Mouiha A., Hafid A., Alaoui A. E. H. Machine Learning and Deep Learning in Computational Finance: A Systematic Review. arXiv. 2025. DOI: https://doi.org/10.48550/arXiv.2511.21588

11. Cohen G. Algorithmic Trading and Financial Forecasting Using Advanced Artificial Intelligence Methodologies. Mathematics. 2022. Vol. 10, No. 18. Article 3302. DOI: https://doi.org/10.3390/math10183302

12. Reis P., Serra A. P., Gama J. The Role of Deep Learning in Financial Asset Management: A Systematic Review. arXiv. 2025. DOI: https://doi.org/10.48550/arXiv.2503.01591

13. Khattak B. H. A., Shafi I., Khan A. S., Flores E., García Lara R. et al. A Systematic Survey of AI Models in Financial Market Forecasting for Profitability Analysis. IEEE Access. 2023. Vol. 11. P. 125359–125380. DOI: https://doi.org/10.1109/ACCESS.2023.3330156

14. Yang H., Liu X.-Y., Wang C. D. FinGPT: Open-Source Financial Large Language Models. arXiv. 2023. DOI: https://doi.org/10.48550/arXiv.2306.06031

15. Kong Y., Nie Y., Dong X., Mulvey J. M., Poor H. V. et al. Large Language Models for Financial and Investment Management: Applications and Benchmarks. The Journal of Portfolio Management. 2024. Vol. 51, No. 2. P. 162–210. DOI: https://doi.org/10.3905/jpm.2024.1.645

16. Jadhav A., Mirza V. Large Language Models in equity markets: applications, techniques, and insights. Frontiers in Artificial Intelligence. 2025. Vol. 8. Article 1608365. DOI: https://doi.org/10.3389/frai.2025.1608365

17. Tatarinov N., Sukhani S., Shah A., Chava S. Language Modeling for the Future of Finance: A Survey into Metrics, Tasks, and Data Opportunities.

arXiv. 2025. DOI: https://doi.org/10.48550/arXiv.2504.07274

18. Divate M., Jadhav P., Jha A., Joshi S., Darak K. Harnessing LLMs for Financial Forecasting: A Systematic Review of Advances in Stock Market Prediction and Portfolio Optimization. International Journal for Research in Applied Science and Engineering Technology. 2024. DOI: https://doi.org/10.22214/ijraset.2024.65283

19. Bhardwaj S. Artificial Intelligence in Wealth Management: Transforming the Future of Financial Advisory Services. Journal of Multidisciplinary Knowledge. 2025. Vol. 5, No. 2. P. 85–96. DOI: https://doi.org/10.36676/jmk.v5.i2.79

20. Tahvildari M. Integrating generative AI in Robo-Advisory: A systematic review of opportunities, challenges, and strategic solutions. Multidisciplinary Reviews. 2025. Vol. 8, No. 12. DOI: https://doi.org/10.31893/multirev.2025379

21. Boddu K., Ram M. V., Kamarajugadda T. V. R., Moorthygari S. L., Bommisetti R. K. The intelligent finance ecosystem: AI applications in banking and fintech for enhanced decision-making. Asian Economic and Financial Review. 2025. Vol. 15, No. 11. P. 1694–1713. DOI: https://doi.org/10.55493/5002.v15i11.5660

22. Chen Y., Zhao C., Xu Y., Nie C., Zhang Y. Year-over-Year Developments in Financial Fraud Detection via Deep Learning: A Systematic Literature Review. arXiv. 2025. DOI: https://doi.org/10.48550/arXiv.2502.00201

23. Chen Y., Zhao C., Xu Y., Nie C., Zhang Y. Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications. Data Science and Management. 2025. DOI: https://doi.org/10.1016/j.dsm.2025.08.002

24. Chen S., Ren S. AI-enabled Forecasting, Risk Assessment, and Strategic Decision Making in Finance. Frontiers in Business and Finance. 2025. Vol. 2, No. 2. P. 274–295. DOI: https://doi.org/10.71465/fbf397

Сергій САВЧЕНКО
Хмельницький національний університет
Херсонський державний університет

ШТУЧНИЙ ІНТЕЛЕКТ У ФІНАНСОВИХ ТЕХНОЛОГІЯХ: МЕТОДИ, ЗАСТОСУВАННЯ ТА СУЧАСНІ ТЕНДЕНЦІЇ

У цьому дослідженні систематизовано сучасні методи генеративного штучного інтелекту, зокрема великих мовних моделей (LLM), та проаналізовано підходи до їх застосування у галузі фінансових технологій. Проведено узагальнення існуючих стратегій використання LLM у FinTech, включаючи zero-shot, few-shot, fine-tuning, Retrieval-Augmented Generation (RAG) та навчання моделей з нуля. Виконано порівняльний аналіз їх вартості та складності впровадження, визначено найбільш відповідні варіанти інтеграції LLM в залежності від типу прикладної задачі. У роботі представлено алгоритм автоматичного ребалансування інвестиційного портфеля, який поєднує класичну портфельну теорію Марковіца (MPT), прогнозування цін фінансових інструментів використовуючи LSTM-мережі та аналіз сигналів технічного аналізу. Запропоновано розширену версію алгоритму ребалансування, яка доповнює традиційні кількісні методи двома LLM-компонентами (модуль аналізу ринкових настроїв та модуль обробки фінансової звітності). Інтеграція зазначених компонентів забезпечує можливість обробки неструктурованих даних, таких як фінансові новини, публікації в соціальних мережах, квартальні та річні фінансові звіти компаній, тощо. Використання таких даних дозволяє суттєво розширити вхідний набір даних моделей прогнозування цін, що може підвищити якість прийнятих інвестиційних рішень. Спираючись на проаналізовані наукові публікації, показано, що поєднання технічних та фундаментальних фінансових показників з оцінкою ринкових настроїв сприяє підвищенню точності прогнозування цін на фінансові інструменти. В роботі обґрунтовано перспективність використання запропонованого методу ребалансування інвестиційного портфелю у автоматизованих системах фінансового консультування (Robo-Advisor). Зазначено основні обмеження дослідження, насамперед, необхідність практичної перевірки алгоритму ребалансування на реальних ринкових даних. Визначено напрями подальших досліджень, пов'язані з експериментальною перевіркою запропонованої моделі на історичних даних різних періодів та подальшу оптимізацію LLM-компонентів спираючись на результати експериментів.

Ключові слова: генеративний штучний інтелект, великі мовні моделі, машинне навчання, фінансові технології, ребалансування інвестиційного портфеля, аналіз ринкових настроїв.