A Survey of Financial AI: Architectures, Advances and Open Challenges

Junhua Liu
Forth AI
j@forth.ai

Abstract

Financial AI empowers sophisticated approaches to financial market forecasting, portfolio optimization, and automated trading. This survey provides a systematic analysis of these developments across three primary dimensions: predictive models that capture complex market dynamics, decision-making frameworks that optimize trading and investment strategies, and knowledge augmentation systems that leverage unstructured financial information. We examine significant innovations including foundation models for financial time series, graph-based architectures for market relationship modeling, and hierarchical frameworks for portfolio optimization. Analysis reveals crucial trade-offs between model sophistication and practical constraints, particularly in high-frequency trading applications. We identify critical gaps and open challenges between theoretical advances and industrial implementation, outlining open challenges and opportunities for improving both model performance and practical applicability[1].

1 Introduction

Recent advancements in artificial intelligence, particularly large language models (LLMs), have significantly transformed quantitative finance. These innovations span multiple domains including predictive modeling, decision making, and knowledge retrieval, enabling more sophisticated approaches to market analysis and trading automation. While Large Language Models (LLMs) have garnered significant attention for their capabilities in natural language processing and reasoning, parallel developments in specialized architectures for financial applications demonstrate equal innovation and practical impact.

Classical approaches to financial modeling face fundamental limitations in capturing complex market dynamics, handling non-stationary distributions, and integrating diverse information sources. Recent work addresses these challenges through three primary directions: architectural innovations in deep learning models, methodological advances in training and optimization, and practical improvements in deployment and scalability. These developments enable more robust prediction under market uncertainty, more efficient portfolio optimization under constraints, and more sophisticated trading strategies incorporating multiple information sources.

1.1 Related Work

Several recent surveys have explored the applications of LLMs in finance from different perspectives. [Lee *et al.*, 2024] present the first comprehensive review of financial LLMs (FinLLMs) including their evolution from general-domain LMs to financial-domain LMs, and provide extensive coverage of model architectures, training data, and benchmarks. [Li *et al.*, 2023] review the current approaches employing LLMs in finance and propose a decision framework to guide their adoption, focusing on practical implementation considerations. [Dong *et al.*, 2024a] provide a scoping review on ChatGPT and related LLMs specifically in accounting and finance, highlighting use cases in these domains. [Zhao *et al.*, 2024] comprehensively explore the integration of LLMs into various financial tasks and applications. [Ding *et al.*, 2024a] focus specifically on LLM agents in financial trading, examining architectures, data inputs, and empirical performance through systematic analysis of 27 relevant papers.

Figure 1 shows the taxonomy of our work, which provides greater breadth across financial applications while maintaining rigorous depth in specific areas, as compared to the current surveys. Undoubtedly, these surveys have advanced our understanding of LLMs in finance. However, there remain significant gaps in addressing the broader implications for financial decision-making and industry practices. Notably, many surveys lack detailed analysis of real-world deployment challenges, integration with existing systems, and comprehensive evaluation frameworks that consider both technical and practical aspects. Our survey addresses these gaps by providing an expanded view of implementation challenges and opportunities.

1.2 Contributions

Our survey makes four primary contributions that distinguish it from prior work:

First, we provide a systematic analysis of the common formulations, techniques, evaluations across the complete spec-

[1]Full list of papers and summary slides are available at: https://github.com/junhua/awesome-finance-ai-papers.

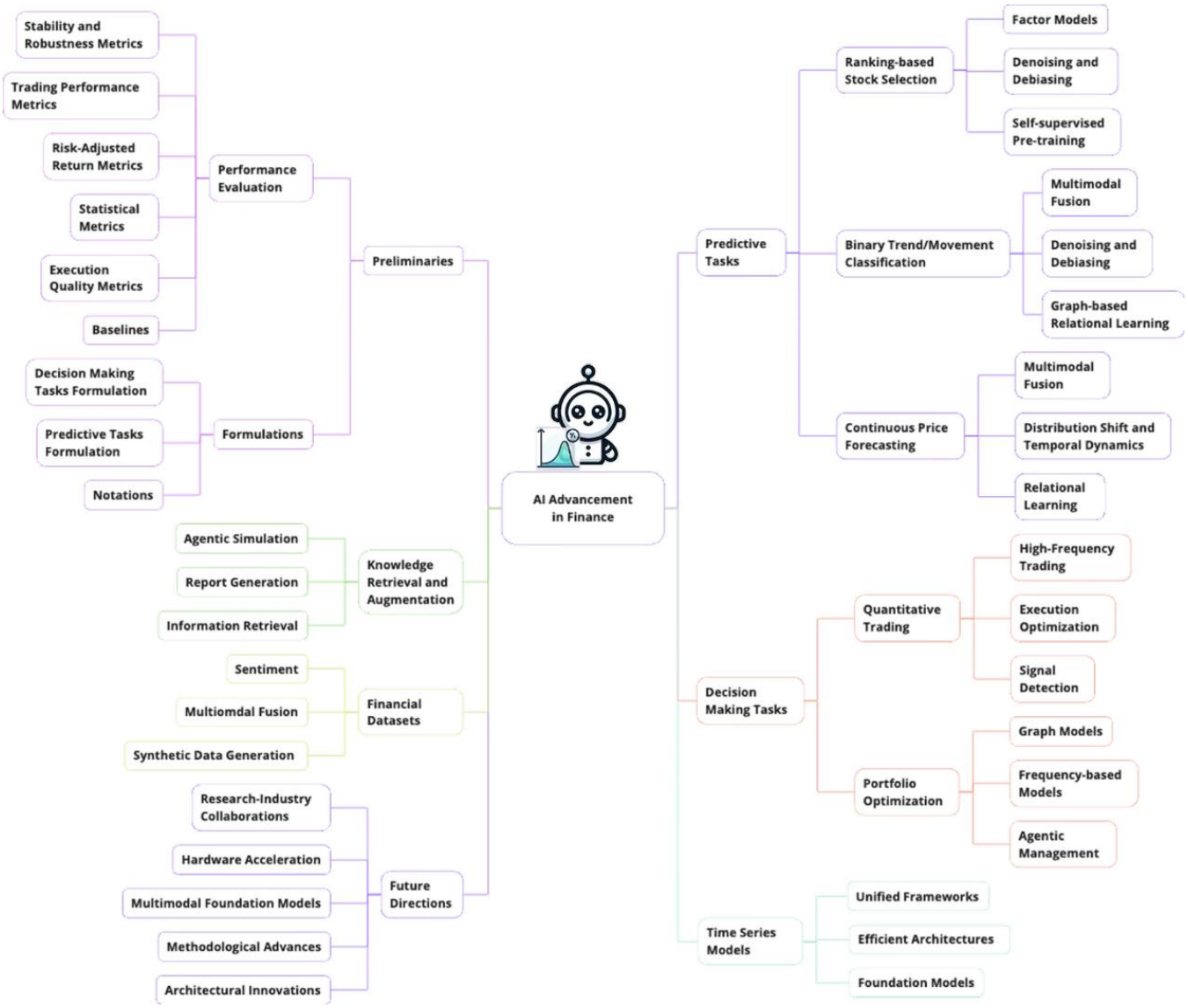

Figure 1: Taxonomy of the survey paper

trum of financial AI applications. The analysis results in a comprehensive task-oriented categorization of the surveyed works.

Second, we present a comprehensive review of architectural and methodological advances beyond LLMs, including wide variations of Graph Neural Networks, Reinforcement Learning, and time series architectures, providing insights for optimal model selection based on specific financial applications.

Third, we identify critical gaps between theories and applications, bridging the gap between academia and industry, and inspiring future research investigation for both researchers and practitioners in advancing the field.

Through rigorous examination of recent work across both academic literature and industry applications, we demonstrate how innovations in AI architecture and methodology enable more sophisticated approaches to financial modeling while highlighting crucial areas for future research. Our analysis emphasizes practical implementation considerations alongside theoretical advancements, providing a uniquely comprehensive resource for researchers and practitioners in computational finance.

1.3 Survey Organization

The survey is organized into eleven core sections. Section 2 establishes the mathematical preliminaries and notations for both predictive and decision-making tasks in financial markets. Section 2.4 presents a comprehensive evaluation framework encompassing statistical accuracy metrics and trading performance measures. Sections 3 to 5 examine predictive tasks: continuous price forecasting, binary trend classification, and ranking-based stock selection, analyzing architectural innovations and empirical validations. Section 6 and 7

investigate decision-making tasks spanning portfolio optimization and quantitative trading, with emphasis on multi-agent frameworks and execution strategies. Section 8 explores knowledge retrieval and augmentation, focusing on automated analysis and market simulation. Section 9 surveys recent financial datasets and benchmarks, while Section 10 examines specialized time series models. Finally, Section 11 identifies promising research directions and open challenges in financial AI. Throughout, we maintain rigorous analysis of theoretical foundations while emphasizing practical deployment considerations.

2 Preliminaries

Financial market applications can be broadly categorized into predictive tasks (continuous price forecasting, binary trend classification, ranking-based selection) and decision-making tasks (portfolio optimization, quantitative trading). Despite their distinct objectives, these tasks share common mathematical foundations and feature spaces. This section presents a unified formulation framework and notations that encompass both prediction and decision-making problems.

2.1 Notation and Problem Setup

Let $\mathcal{S} = \{s_1, ..., s_N\}$ denote a pool of N assets. For each asset s_i, we observe a temporal sequence of T historical features $\mathbf{X}_i = \{\mathbf{x}_i^t\}_{t=1}^T \in \mathbb{R}^{T\times d}$, with d being the feature dimension. The feature vector $\mathbf{x}_i^t$ encompasses three categories of market information. The first category consists of price-derived features, including open, high, low, close prices, and trading volume, which capture direct market activities. The second category comprises technical indicators such as moving averages, relative strength index (RSI), and Bollinger bands, which provide insights into market momentum and trends. The third category includes fundamental metrics like price-to-earnings ratio, market capitalization, and liquidity measures, which reflect underlying asset value and market characteristics.

The formulation incorporates several components to capture broader market dynamics. The market state $\mathbf{M}^t \in \mathbb{R}^{d_m}$ represents broader market conditions, economic indicators, and risk factors. The relational structure $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ models complex relationships between assets, where $\mathcal{V}$ represents the set of assets, $\mathcal{E}$ captures their interactions, and $\mathcal{R}$ defines different types of relationships. Textual information $\mathcal{T}_i^t$ from sources such as news and financial reports provides qualitative context. The agent's state $\mathbf{s}_t$ includes portfolio positions $\mathbf{w}_t \in \mathbb{R}^N$, capital B_t, and transaction costs $\mathbf{c}_t$, essential for decision-making tasks.

2.2 Predictive Tasks Formulations

Continuous Price Forecasting

The continuous forecasting task learns a function f_θ mapping historical observations to future values over horizon h:

$$\hat{\mathbf{y}}_i^{t+1:t+h} = f_\theta(\mathbf{x}_i^{t-w+1:t}, \mathbf{M}^t, \mathcal{G}, \mathcal{T}_i^t) \tag{1}$$

The target variable $\mathbf{y}$ can take several forms: raw price levels $y_i^t = p_i^t$, returns $y_i^t = (p_i^t - p_i^{t-1})/p_i^{t-1}$, or volatility $y_i^t = \sqrt{\mathbb{E}[(r_i^t - \mu_i^t)^2]}$, where r_i^t represents the log-return and μ_i^t its mean.

Table 1: Summary of Notations in Problem Formulations

Symbol	Description
$\mathcal{S}$	Asset pool
$\mathbf{x}_i^t$	Feature vector for asset i at time t
$\mathbf{X}_i$	Temporal feature sequence for asset i
$\mathbf{M}^t$	Market state at time t
$\mathcal{G}$	Asset relationship graph
$\mathcal{T}_i^t$	Textual information for asset i at time t
$\mathbf{w}_t$	Portfolio weights at time t
B_t	Available capital at time t
$\mathbf{c}_t$	Transaction costs at time t
f_θ	Prediction function with parameters θ
π_ϕ	Policy function with parameters ϕ
h_ψ	Encoder function with parameters ψ
p_i^t	Price of asset i at time t
r_i^t	Return of asset i at time t
δ	Threshold for binary classification

Binary Trend Classification

The classification task extends the general forecasting framework by discretizing price movements into directional categories. The model learns a mapping function that predicts movement direction:

$$\hat{y}_i^{t+h} = f_\theta(\mathbf{x}_i^{t-w+1:t}, \mathbf{M}^t, \mathcal{G}, \mathcal{T}_i^t) \tag{2}$$

with target label:

$$y_i^{t+h} = \begin{cases} 1 & \text{if } \frac{p_i^{t+h} - p_i^t}{p_i^t} > \delta \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where δ represents a threshold parameter accounting for transaction costs and market impact.

Ranking-based Selection

The ranking task focuses on learning relative ordering of assets based on their expected future performance through a scoring function:

$$\hat{r}_i^{t+h} = f_\theta(\mathbf{x}_i^{t-w+1:t}, \mathbf{M}^t, \mathcal{G}, \mathcal{T}_i^t) \tag{4}$$

which induces ranking:

$$\pi_t = \text{argsort}(\{\hat{r}_i^{t+h}\}_{i=1}^N) \tag{5}$$

2.3 Decision Making Tasks Formulation

Portfolio optimization and quantitative trading can be formulated as sequential decision-making problems under uncertainty, typically modeled through Markov Decision Processes (MDPs) or their variants.

Portfolio Optimization

The portfolio optimization task aims to determine optimal asset allocations over time. Let $\mathbf{w}_t = [w_1^t, ..., w_N^t]$ represent portfolio weights at time t, where w_i^t denotes the proportion of capital allocated to asset i. The state space $s_t \in \mathcal{S}$ encompasses market features $\mathbf{X}_t$, current portfolio weights $\mathbf{w}_{t-1}$, and available capital B_t:

$$s_t = [\mathbf{X}_t, \mathbf{w}_{t-1}, B_t] \tag{6}$$

The action space $a_t \in \mathcal{A}$ defines target portfolio weights:

$$a_t = \mathbf{w}_t, \quad \text{s.t.} \sum_{i=1}^{N} w_i^t = 1, \quad w_i^t \geq 0 \tag{7}$$

The reward function incorporates both returns and transaction costs:

$$r_t = \sum_{i=1}^{N} w_i^t r_i^t - c \sum_{i=1}^{N} |w_i^t - w_i^{t-1}| \tag{8}$$

The objective is to learn a policy $\pi_\theta : \mathcal{S} \to \mathcal{A}$ maximizing expected cumulative returns:

$$\max_\theta \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t r_t | \pi_\theta\right] \tag{9}$$

Quantitative Trading

Trading can be formulated as a POMDP where the true market state is partially observable. The agent's belief state b_t combines observable market features with internal estimates of latent variables:

$$b_t = [\mathbf{X}_t, \mathbf{h}_t, p_t, v_t] \tag{10}$$

where $\mathbf{h}_t$ represents hidden state estimates, p_t is position size, and v_t is remaining capital.

The action space includes both trade direction and size:

$$a_t = (d_t, q_t), \quad d_t \in \{-1, 0, 1\}, \quad 0 \leq q_t \leq Q_{max} \tag{11}$$

The transition function incorporates market impact:

$$p_{t+1} = p_t + d_t q_t \tag{12}$$

$$v_{t+1} = v_t - d_t q_t (p_t + \alpha q_t) \tag{13}$$

where α models price impact. The reward balances profit against risk:

$$r_t = d_t q_t (p_{t+1} - p_t) - c|q_t| - \lambda \text{Risk}(p_t, q_t) \tag{14}$$

Both tasks can be solved through various reinforcement learning approaches, with policy gradient methods being particularly suitable due to their ability to handle continuous action spaces and complex constraints. The choice between model-based and model-free approaches depends on the trade-off between sample efficiency and computational complexity.

2.4 Performance Evaluation

Definitions and Notations

The following notations are used throughout this section:

Table 2: Summary of Notations In Performance Evaluation

Symbol	Description
R_p	Portfolio return
R_f	Risk-free rate (3-month Treasury bill yield)
R_m	Market return (S&P 500 index)
R_b	Benchmark return (relevant market index)
σ_p	Portfolio volatility
σ_{p-b}	Relative portfolio volatility
σ_{strategy}	Strategy volatility
$\sigma_{\text{benchmark}}$	Benchmark volatility
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
$\mathbb{E}$	Expected value operator
w_i^t	Weight of asset i at time t
q_t	Trading quantity at time t
P_t^e	Expected execution price at time t
P_t^a	Actual execution price at time t

Statistical Metrics

For continuous price forecasting, prediction quality is measured through:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{15}$$

$$\text{IC} = \text{corr}(\hat{\mathbf{y}}, \mathbf{y}) \tag{16}$$

Binary classification employs:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{17}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

Risk-Adjusted Return Metrics

Portfolio performance is assessed through:

$$\text{SR} = \frac{R_p - R_f}{\sigma_p} \tag{19}$$

$$\text{CR} = \frac{R_p - R_f}{\text{MDD}} \tag{20}$$

where Maximum Drawdown (MDD) captures downside risk:

$$\text{MDD} = \max_t \{\max_{\tau \leq t}(V_\tau) - V_t\} / \max_{\tau \leq t}(V_\tau) \tag{21}$$

Alpha measures excess returns over a benchmark:

$$\alpha = R_p - [\beta(R_m - R_f) + R_f] \tag{22}$$

The Information Ratio quantifies consistency of excess returns:

$$\text{IR} = \frac{R_p - R_b}{\sigma_{p-b}} \tag{23}$$

Long-term performance is captured through:

$$\text{AR} = \frac{252}{T} \sum_{t=1}^{T} r_t \tag{24}$$

$$\text{CAGR} = (1 + R_T)^{252/T} - 1 \tag{25}$$

Volatility-adjusted measures include:

$$\lambda_{\text{Vol}} = \frac{\sigma_{\text{strategy}}}{\sigma_{\text{benchmark}}} \tag{26}$$

Trading Performance Metrics
Trading strategy evaluation incorporates:

$$\text{Net Return} = \sum_{t=1}^{T} \left(r_t - c|q_t| - \alpha q_t^2 \right) \tag{27}$$

The Win-Loss Ratio measures trading effectiveness:

$$\text{WLR} = \frac{\text{Number of Profitable Trades}}{\text{Number of Loss-Making Trades}} \tag{28}$$

Execution Quality Metrics
Portfolio optimization strategies are evaluated on rebalancing efficiency:

$$\text{Turnover} = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} |w_i^t - w_i^{t-1}| \tag{29}$$

Trading execution quality is measured through Implementation Shortfall:

$$\text{IS} = \frac{1}{T} \sum_{t=1}^{T} (P_t^e - P_t^a) q_t \tag{30}$$

where P_t^e and P_t^a are expected and actual execution prices.

Stability and Robustness Metrics
Strategy stability is assessed through portfolio diversification measures:

$$\text{Effective N} = \frac{1}{\sum_{i=1}^{N} w_i^2}, \quad \text{HHI} = \sum_{i=1}^{N} w_i^2 \tag{31}$$

Robustness across market regimes is evaluated through conditional metrics:

$$\text{Regime SR} = \frac{R_p^k - R_f}{\sigma_p^k}, \quad k \in \{\text{bull}, \text{bear}, \text{neutral}\} \tag{32}$$

Strategy timing ability is assessed through the Treynor-Mazuy model:

$$R_p - R_f = \alpha + \beta(R_m - R_f) + \gamma(R_m - R_f)^2 + \epsilon \tag{33}$$

Multi-factor performance attribution employs the Fama-French framework:

$$R_p - R_f = \alpha + \beta_M(R_m - R_f) + \beta_S \text{SMB} + \beta_H \text{HML} + \beta_M \text{MOM} + \epsilon \tag{34}$$

Baselines
Performance is benchmarked against:

- Traditional strategies: Equal-weight, Minimum Variance, Maximum Sharpe
- Market indices: S&P 500, NASDAQ Composite
- Statistical models: ARIMA, GARCH
- Machine learning baselines: LSTM, GRU, Transformer

All metrics are computed using walk-forward optimization with appropriate lookback windows to prevent forward-looking bias.

3 Continuous Price Forecasting

Continuous price forecasting represents a fundamental challenge in quantitative finance, combining elements of time series analysis, market microstructure, and behavioral finance. Recent works in AI have led to tremendous improvements in forecasting accuracy and robustness, particularly through the integration of complex market relationships, adaptive learning mechanisms, and multimodal data sources. As summarized in Table 3, these works center around three primary areas, namely, relational learning approaches that capture market structure, distribution shift modeling that addresses temporal dynamics, and hybrid approaches that integrate multiple information modalities. The contributions range from theoretical innovations in model architecture (e.g., MASTER's dynamic correlation modeling) to practical applications in performance metrics (e.g., DIFFSTOCK's improvements in Sharpe ratios). This section examines these developments systematically, beginning with a rigorous mathematical formulation of the forecasting problem, followed by detailed analysis of methodological innovations and their empirical validation across different market contexts.

3.1 Relational Learning Approaches

The complex interdependencies in financial markets naturally motivate graph-based and relational learning approaches, which can capture both explicit market relationships and implicit correlations between assets. Recent works have demonstrated the effectiveness of such approaches through increasingly sophisticated architectures. MDGNN [Xu *et al.*, 2024] established a fundamental framework by modeling multi-relational market structure through a hierarchical graph representation. The framework employs a transformer structure with ALIBI position encoding to capture temporal evolution patterns, demonstrating superior performance on China's CSI300 index with significant improvements in Information Coefficient and Cumulative Return metrics.

Extending this relational modeling paradigm, MASTER [Li *et al.*, 2024b] introduced a more dynamic approach through a specialized transformer architecture. Unlike MDGNN's static graph structure, MASTER alternates between intra-stock and inter-stock information aggregation, enabling the capture of time-varying relationships. Its market-guided gating mechanism dynamically selects relevant features based on market conditions, achieving a 13% improvement in ranking metrics and 47% in portfolio-based metrics on CSI300 and CSI800 markets.

DANSMP [Zhao *et al.*, 2022] further advanced the field by incorporating higher-order market relationships through a comprehensive market knowledge graph (MKG). The model expands beyond traditional asset-to-asset relationships by integrating executive entities and implicit connections, processing these diverse signals through dual attention mechanisms. This richer relational modeling led to superior performance on the CSI300E dataset, achieving investment returns of 16.97% with Sharpe ratios of 4.628, demonstrating the value of incorporating executive-level relationships in prediction models.

Table 3: Summary of Recent Contributions in Continuous Price Forecasting

Work	Key Innovation	Methodology	Primary Results
MDGNN [Xu *et al.*, 2024]	Multi-relational graph incorporating industry and institutional relationships	Hierarchical graph embedding with ALIBI position encoding	IC improvement: 15% (CSI300) Stronger performance on larger datasets
MASTER [Li *et al.*, 2024b]	Dynamic stock correlation modeling through alternating attention	Market-guided gating mechanism for feature selection	Ranking metrics: +13% Portfolio metrics: +47% (CSI300, CSI800)
DoubleAdapt [Zhao *et al.*, 2023]	Dual adaptation for distribution shifts in streaming data	Multi-head transformation layers for feature and label adaptation	Consistent improvement across LSTM, GRU, Transformer implementations Model-agnostic adaptability
DPA-STIFormer [Yan and Tan, 2024]	Feature-centric temporal modeling	Double-path mechanism with adaptive gating	Superior performance across CSI500, CSI1000, NASDAQ, NYSE Improved IC and Sharpe Ratio
DANSMP [Zhao *et al.*, 2022]	Executive-level market knowledge graph	Dual attention network for complex entity interactions	Returns: 16.97% Sharpe Ratio: 4.628 (CSI300E)
GINN [Xu *et al.*, 2024]	Integration of GARCH theory with neural networks	GARCH-based regularization in LSTM loss function	Consistent outperformance across 7 global indices Strong zero-shot performance
DIFFSTOCK [Daiya *et al.*, 2024]	Diffusion models for market prediction	Adaptive noise scheduling for stock volatility	Sharpe Ratio improvement: NYSE: +7.92% NASDAQ: +6.18%
[Wang *et al.*, 2024b]	LLM integration for news processing	Text-based transformation of numerical forecasting	Superior performance in event-driven market shifts Multi-domain applicability

3.2 Distribution Shift and Temporal Dynamics

The non-stationary nature of financial markets presents a fundamental challenge to traditional machine learning approaches, spurring the development of methods specifically designed to handle temporal distribution shifts. DoubleAdapt [Zhao *et al.*, 2023] addressed this challenge through a meta-learning framework that implements dual adaptation mechanisms. The framework's multi-head transformation layers adapt both features and labels into locally stationary distributions, while its model adapter learns initialization parameters that enable quick adaptation to new data. This approach showed consistent improvements across various time-series architectures, effectively handling both gradual and sudden market shifts.

Building on these insights, DPA-STIFormer [Yan and Tan, 2024] introduced a novel perspective on temporal modeling by treating features rather than time steps as tokens. Its double-path mechanism adaptively learns stock relationships, while a specialized decoder decomposes predictions into mean and deviation components. This architecture proved particularly effective across diverse market conditions, demonstrating robust performance on four major markets (CSI500, CSI1000, NASDAQ, NYSE).

DIFFSTOCK [Daiya *et al.*, 2024] approached the distribution shift challenge from a generative modeling perspective, leveraging denoising diffusion models. The framework's Masked Relational Transformer architecture processes different relationship types through separate attention heads, while its adaptive noise schedule accounts for both individual stock volatility and intra-cluster dynamics. This approach greatly improves Sharpe ratios (7.92% and 6.18% increases for NYSE and NASDAQ respectively), though with increased computational requirements.

3.3 multimodal Fusion

The integration of traditional financial theory with modern machine learning techniques has emerged as a promising direction for improving forecast robustness. GINN [Xu *et al.*, 2024] exemplifies this approach by bridging classical GARCH models with neural networks. The framework incorporates GARCH predictions as a regularization term in the LSTM's loss function, effectively combining statistical finance theory with deep learning flexibility. Evaluated on seven global market indices over a 30-year period, both GINN and its variant GINN-0 demonstrated consistent outperformance over traditional approaches, though with a trade-off in capturing market volatility.

Taking a different approach to multimodal integration, [Wang *et al.*, 2024b] leveraged recent works in large language models to incorporate unstructured news data into forecasting. The framework transforms numerical forecasting into a text-based task, employing LLM-based agents for news filtering and prediction evaluation. This approach proved particularly effective in capturing sudden market shifts caused by external events, demonstrating the value of incorporating qualitative information in traditionally quantitative predictions.

Table 4: Comparison of Recent Continuous Price Forecasting Approaches

Model	Temporal Modeling	Relational Learning	Distribution Shift	multimodal Integration	Compute Cost	Market Coverage (Datasets)
MDGNN	✓	✓	-	-	Medium	China (CSI100, CSI300)
MASTER	✓	✓	✓	-	High	China (CSI300, CSI800)
DoubleAdapt	✓	-	✓	-	Low	China (CSI300, CSI500)
DPA-STIFormer	✓	✓	✓	-	High	China (CSI500, CSI1000), US (NASDAQ, NYSE)
GINN	✓	-	-	✓	Low	Global (7 major market indices, 30-year period)
DIFFSTOCK	✓	✓	✓	-	Very High	US (NASDAQ, NYSE, StockNet)
DANSMP	✓	✓	-	✓	High	China (CSI100E, CSI300E)

3.4 Comparative Analysis

The reviewed approaches exhibit distinct characteristics in terms of their modeling capacity, computational efficiency, and practical applicability. Table 4 summarizes the key aspects of representative models.

In terms of architectural design, models exhibit a clear trade-off between modeling capacity and computational efficiency. MASTER and DPA-STIFormer achieve superior performance through sophisticated attention mechanisms but require significant computational resources. In contrast, DoubleAdapt and GINN maintain efficiency through focused adaptation mechanisms and classical model integration, respectively.

Market coverage and generalization capabilities also vary significantly. While models like MDGNN and DANSMP demonstrate strong performance on Chinese markets, DPA-STIFormer and DIFFSTOCK show broader applicability across both US and Chinese markets. GINN stands out with its global market coverage, though focusing on major indices rather than individual stocks. This market specialization versus generalization represents a key consideration in model selection and development.

The integration of domain knowledge presents another key differentiator. GINN's incorporation of GARCH theory leads to more interpretable and theoretically grounded predictions, while purely data-driven approaches like DIFFSTOCK achieve higher performance at the cost of reduced interpretability.

4 Binary Trend/Movement Classification

Binary trend classification represents a critical task in quantitative finance that focuses on predicting directional price movements in financial markets. Unlike continuous price forecasting which aims to predict exact values, this task addresses the fundamental question of price movement direction, making it particularly relevant for trading strategies and risk management. Prediction accuracy are improved by market structure modeling, noise handling, and information fusion. These developments span three primary directions: graph-based relational modeling that captures complex market dependencies, denoising techniques that address market noise and distribution shifts, and multimodal approaches that integrate diverse information sources, as summarized in Table 5.

4.1 Graph-based Relational Learning

The complex interdependencies in financial markets have motivated the development of sophisticated graph-based architectures that capture both explicit and implicit relationships between assets. ECHO-GL [Liu *et al.*, 2024a] established a comprehensive framework for modeling market relationships through heterogeneous graph learning. The model introduces a dual-mechanism approach combining spatial relational modeling with stock dynamics modeling. The spatial component aggregates multimodal information across the graph structure, while the dynamics component captures post-earnings announcement effects through learnable stochastic processes. When evaluated on the S&P 500 and NASDAQ-100 indices from 2018-2023, this architecture achieved a 2.297% increase in F1 score and 15.629% increase in MCC over baseline methods, with particularly strong performance during earnings seasons. MGDPR [You *et al.*, 2024] advanced the field by introducing dynamic relationship modeling through information entropy and signal energy. The framework's innovation lies in its multi-relational diffusion process, which adaptively learns and updates relationship strengths between assets. Implementation challenges, particularly in computational efficiency for large-scale markets, were addressed through sparse matrix operations and optimized graph convolutions. This approach enabled more effective capture of market dynamics, leading to consistent outperformance in next-day trend forecasting across NASDAQ, NYSE, and Shanghai markets over a seven-year test period from 2016-2023.

4.2 Denoising and Debiasing

Financial markets are characterized by high noise levels and non-stationary distributions, leading to the development of specialized denoising and adaptation techniques. LARA [Zeng *et al.*, 2024] introduced a comprehensive framework combining locality-aware attention with iterative refinement. The framework's two-stage approach first identifies profitable trading opportunities through metric learning, then iteratively refines noisy labels to improve prediction robustness. Evaluated on high-frequency data from China's A-share market (2020-2023), cryptocurrency markets (2021-2023), and ETF markets (2019-2023), this architecture demonstrated remarkable resilience in volatile market conditions, achieving up to 59.1% precision while maintaining computational efficiency through optimized attention mechanisms. MANA-Net [Wang and Ma, 2024] tackled the critical challenge of "aggregated sentiment homogenization" in financial news analysis. The model employs a dynamic market-news attention mechanism that weights news items based on their market relevance, integrating news aggregation and market prediction into a unified framework. Validated on an extensive dataset spanning 2003-2018 with over 2.7 million news items from major financial news sources, MANA-Net achieved a 1.1% increase in Profit and Loss and a 0.252 increase in daily Sharpe ratio, demonstrating the value of sophisticated news aggregation in market prediction. These binary trend classification approaches exhibit distinct characteristics in terms of their modeling capacity, computational efficiency, and practical applicability. Table 6 summarizes the key aspects of representative models.

Table 5: Summary of Recent Contributions in Binary Trend Classification

Work	**Key Innovation**	**Methodology**	**Primary Results**
ECHO-GL [Liu *et al.*, 2024a]	Heterogeneous graph learning from earnings calls data	Dual-mechanism approach combining spatial relations with stock dynamics	F1: +2.297% MCC: +15.629% (S&P500, NASDAQ-100)
MGDPR [You *et al.*, 2024]	Dynamic relationship modeling through information entropy	Multi-relational diffusion process with parallel retention	Consistent outperformance across NASDAQ, NYSE, Shanghai (2016-2023)
LARA [Zeng *et al.*, 2024]	Two-stage denoising framework with locality-aware attention	Iterative refinement labeling for noise reduction	Precision: 59.1% Improved win-loss ratio in volatile markets
MANA-Net [Wang and Ma, 2024]	Dynamic market-news attention mechanism	Trainable sentiment aggregation optimized for prediction	P&L: +1.1% Daily SR: +0.252
SH-Mix [Jain *et al.*, 2024]	Hierarchical multimodal augmentation strategy	Modality-specific and span-based mixing techniques	Performance improvement: 3-7% across tasks
SEP [Koa *et al.*, 2024a]	Self-reflective LLM framework for autonomous learning	Three-stage pipeline with PPO optimization	Superior performance in both prediction and explanation quality

Table 6: Comparison of Recent Binary Trend Classification Approaches

Model	Temporal Modeling	Graph Learning	Distribution Handling	Multimodal Integration	Compute Cost	Market Coverage (Datasets)
ECHO-GL	✓	✓	-	✓	High	US (S&P500, NASDAQ-100)
MGDPR	✓	✓	✓	-	Medium	Global (NASDAQ, NYSE, SSE)
LARA	✓	-	✓	-	Low	China (A-share), Crypto, ETFs
MANA-Net	✓	-	✓	✓	Medium	US (S&P500)
SH-Mix	✓	-	-	✓	Medium	US (S&P500)
SEP	-	-	-	✓	Very High	Global (Multiple Indices)

4.3 Multimodal Fusion

The integration of multiple data modalities presents unique challenges in feature alignment and data scarcity. SH-Mix [Jain *et al.*, 2024] developed a hierarchical augmentation strategy operating at both local and global levels. The framework performs modality-specific mixing based on feature importance locally, while applying span-based mixing on fused representations globally. Built upon an attention-driven fusion architecture and evaluated on earnings call datasets from S&P 500 companies (2019-2023), the approach achieved 3-7% improvement over existing methods while demonstrating strong generalization capabilities across various multimodal tasks. The emergence of large language models has enabled new approaches to feature extraction and prediction. SEP [Koa *et al.*, 2024a] implements a three-stage pipeline combining summarization, explanation, and prediction components. The framework, tested on market data and financial news from 2020-2023, allows LLMs to autonomously learn and improve stock predictions without human expert intervention, demonstrating superior performance over both traditional deep learning and existing LLM approaches in both prediction accuracy and explanation quality.

4.4 Comparative Analysis

The architectural approaches demonstrate clear trade-offs between modeling sophistication and computational efficiency. Graph-based models like ECHO-GL and MGDPR achieve superior prediction accuracy through comprehensive market structure modeling but require significant computational resources. In contrast, LARA maintains efficiency through focused denoising mechanisms while sacrificing some modeling capacity.

Market coverage and generalization capabilities vary significantly across approaches. MGDPR demonstrates strong cross-market applicability spanning US, Chinese, and European markets, while models like MANA-Net and SH-Mix show specialized performance on US markets. SEP's LLM-based approach offers global coverage but with higher computational requirements and less predictable performance characteristics.

The integration of domain knowledge presents another key differentiator. ECHO-GL's incorporation of earnings call information and MANA-Net's sophisticated news processing lead to more interpretable predictions, while purely data-driven approaches like LARA achieve robust performance through statistical learning. This trade-off between interpretability and performance represents a crucial consideration in model selection.

5 Ranking-based Stock Selection

Ranking-based stock selection represents a fundamental task in quantitative finance that aims to order a universe of assets based on their expected performance. Unlike continuous price forecasting which predicts exact values or binary classification which focuses

Table 7: Summary of Recent Contributions in Ranking-based Stock Selection

Work	**Key Innovation**	**Methodological Contribution**	**Primary Results**
CI-STHPAN [Xia *et al.*, 2024b]	Channel-independent pre-training with dynamic hypergraph learning	Self-supervised framework with reversible normalization	IR: +21.3%, IC: +15.7%, Outperforms SOTA on NYSE, NASDAQ
ADB-TRM [Chen *et al.*, 2024]	Meta-learning framework for dual-level bias mitigation	Temporal-relational adversarial training with global distribution adaptation	Returns: +28.41% Risk-adjusted Returns: +9.53% (NYSE, NASDAQ, TSE)
RSAP-DFM [Xiang *et al.*, 2024]	Dual regime-shifting mechanism for factor modeling	Gradient-based posterior factors with adversarial learning	Factor Returns: +18.2% Robust macro-state adaptation (A-share market)
RT-GCN [Zheng *et al.*, 2023]	Pure convolutional temporal-relational modeling	Three-strategy information propagation with time-sensitive weighting	Returns: +40.4% Training Time: 13.4× faster (NASDAQ, NYSE, CSI)

Table 8: Comparison of Recent Ranking-based Stock Selection Approaches

Model	Pre-training Integration	Factor Modeling	Distribution Adaptation	Compute Cost	Market Coverage (Datasets)
CI-STHPAN	✓	-	✓	High	US (NYSE, NASDAQ)
RT-GCN	-	-	✓	Low	Global (NYSE, NASDAQ, CSI)
ADB-TRM	-	-	✓	Medium	Global (NYSE, NASDAQ, TSE)
RSAP-DFM	-	✓	✓	High	China (A-share)

on directional movement, ranking approaches learn relative orderings that directly inform portfolio construction decisions. Ranking accuracy and robustness are optimized through innovations in self-supervised learning, bias mitigation, and integration with factor models. As summarized in Table 7, these works provide three primary directions: pre-training approaches that leverage unlabeled data, debiasing techniques that address various market biases, and hybrid approaches that combine deep learning with traditional finance theory.

5.1 Self-supervised Pre-training

The abundance of unlabeled financial data has motivated the development of sophisticated pre-training approaches that learn meaningful representations before fine-tuning for ranking tasks. CI-STHPAN [Xia *et al.*, 2024b] established a comprehensive framework combining channel-independent processing with dynamic hypergraph learning. The model's innovation lies in constructing adaptive hypergraphs based on time series similarities using Dynamic Time Warping, moving beyond predefined relationships. To address distribution shifts, the framework incorporates reversible instance normalization and employs a two-stage training process. When evaluated on NASDAQ and NYSE markets over five years, this architecture achieved 21.3% improvement in Information Ratio and 15.7% in Information Coefficient, with particularly strong performance in capturing complex market dependencies.

RT-GCN [Zheng *et al.*, 2023] advanced the field through a pure convolutional approach to temporal-relational modeling. The framework's key innovation lies in its unified treatment of temporal patterns and stock relationships through a relation-temporal graph structure. The model employs three relation-aware strategies for information propagation - uniform, weighted, and time-sensitive - with adaptive weighting based on temporal dynamics. This architecture enables significantly faster training while maintaining high accuracy, achieving up to 40.4% improvement in investment returns while reducing computational time by 13.4× compared to existing methods across NASDAQ, NYSE, and CSI markets.

5.2 Denoising and Debiasing

Financial markets exhibit various biases and non-stationary distributions that challenge traditional learning approaches. ADB-TRM [Chen *et al.*, 2024] introduced a comprehensive framework addressing both micro-level biases in stock data and macro-level distribution shifts. The model employs temporal adversarial training to handle inherent noise while using relational adversarial training to mitigate momentum spillover effects. Its meta-learning framework enables adaptation to changing market conditions through invariant feature extraction. Evaluated across NYSE, NASDAQ, and Tokyo Stock Exchange, this approach demonstrated noticeable improvements of 28.41% in cumulative returns and 9.53% in risk-adjusted returns while maintaining computational efficiency.

5.3 Factor Models

The incorporation of traditional financial theory with modern machine learning has emerged as a promising direction for improving ranking robustness. RSAP-DFM [Xiang *et al.*, 2024] pioneered this approach through a dual regime-shifting mechanism that continuously captures macroeconomic states and their impact on factor dynamics. The framework employs multi-head attention for dynamic factor generation while introducing gradient-based posterior factors through adversarial learning. A novel bilevel optimization algorithm separates factor construction from model optimization, enabling more efficient training. Tested on China's A-share market, the model achieved 18.2% improvement in factor returns while providing explicit interpretability through macroeconomic state mapping.

5.4 Comparative Analysis

These ranking approaches exhibit distinct characteristics in terms of their modeling capacity, computational efficiency, and practical applicability. Table 8 summarizes key aspects of representative models.

The architectural approaches demonstrate clear trade-offs between modeling sophistication and computational efficiency. Pretraining models like CI-STHPAN achieve superior ranking accuracy through comprehensive market structure modeling but require significant computational resources. In contrast, RT-GCN maintains efficiency through focused convolutional processing while achieving competitive performance.

Market coverage and generalization capabilities vary significantly across approaches. RT-GCN and ADB-TRM demonstrate strong cross-market applicability spanning multiple global markets, while RSAP-DFM shows specialized performance on Chinese markets through its integration with factor models.

6 Portfolio Optimization

Portfolio optimization represents a fundamental challenge in quantitative finance that involves allocating capital across multiple assets to maximize risk-adjusted returns. Recent work in Financial AI have significantly improved portfolio management through innovations in multi-agent systems, frequency domain analysis, and risk modeling. As summarized in Table 9, these works centered around three primary directions: agent-based approaches that enable adaptive management, frequency-based methods that capture market dynamics across different time scales, and network-centric techniques that model complex dependencies for risk management.

6.1 Comparative Analysis

These portfolio optimization approaches exhibit distinct characteristics in terms of their modeling capacity, computational efficiency, and practical applicability. Table 10 summarizes key aspects of representative models.

6.2 Agentic Management

The complexity of portfolio management has motivated the development of sophisticated multi-agent architectures that decompose the optimization problem into specialized components. MASA [Li *et al.*, 2024e] established a comprehensive framework using three cooperative agents: a trend observer for market monitoring, a return optimizer for portfolio maximization, and a risk manager for risk minimization. The framework's key innovation lies in its reward-based guiding mechanism that combines return and action rewards to maintain strategy diversity while adapting to market conditions. When evaluated on major indices (CSI300, DJIA, S&P500), this architecture demonstrated superior performance in both stable and volatile markets.

EarnMore [Zhang *et al.*, 2024d] advanced the field through a customizable approach to portfolio management. The framework introduces a maskable token system to represent unfavorable stocks and employs self-supervised learning for relationship modeling. Its one-shot training capability enables efficient adaptation to different stock pools while maintaining performance, achieving 40% profit improvement across diverse market conditions.

6.3 Frequency-based Models

The multi-scale nature of market patterns has inspired approaches that explicitly model different frequency components. FreQuant [Jeon *et al.*, 2024] pioneered this direction through a reinforcement learning framework operating in the frequency domain. The model employs Discrete Fourier Transform to identify market patterns at different frequencies, processing these through a Frequency State Encoder for multi-granular asset representation. This approach achieved up to 2.1× improvement in Annualized Rate of Return across U.S., Korean, and cryptocurrency markets, with particular strength in handling market regime shifts.

TrendTrader [Ding *et al.*, 2024b] complemented frequency analysis with multimodal integration, combining price patterns with news sentiment through a spatial-temporal backbone network. Its incremental learning approach for portfolio weights demonstrated robust performance across DJIA and SSE-50 indices, particularly in capturing sentiment-driven market movements.

6.4 Graph Models

The complex dependencies in financial markets have motivated network-based approaches to risk management. [Hui and Wang, 2024] developed a framework using Graph Theory and Extreme Value Theory to mitigate extremal risks. The approach constructs graphs based on extremal dependencies between stock returns, using maximum independent sets for diversification. Tested on CSI 300 components, this method outperformed traditional sector-based approaches, especially during market downturns.

[Yamagata and Ono, 2024] advanced this direction by replacing sector-based diversification with market graph-based clustering. The framework incorporates turnover sparsity regularization to manage transaction costs while ensuring cluster-based diversification. Experiments on S&P500 demonstrated superior Sharpe ratios compared to traditional methods, particularly in challenging market conditions.

The architectural approaches demonstrate clear trade-offs between modeling sophistication and computational efficiency. Multi-agent systems like MASA achieve superior performance through comprehensive market modeling but require significant computational resources. In contrast, EarnMore maintains efficiency through focused token-based processing while achieving competitive performance.

7 Quantitative Trading

Quantitative trading represents a systematic approach to financial markets that leverages mathematical models and computational methods to develop and execute trading strategies. Recent works in Financial AI have significantly enhanced quantitative trading through innovations in three key areas: strategy development incorporating predictive signals and market dynamics, execution optimization through hierarchical control, and high-frequency trading systems that adapt to market microstructure. Unlike traditional approaches that rely on fixed rules or simple statistical arbitrage, modern quantitative trading systems employ sophisticated deep learning architectures to capture complex market patterns while maintaining computational efficiency for real-time deployment. As summarized in Table 11, these works propose predictive modeling frameworks that extract meaningful signals from noisy market data, execution optimization methods that decompose complex trading decisions into manageable components, and adaptive systems that handle the unique challenges of high-frequency market environments. This section examines these developments systematically, analyzing their theoretical foundations, architectural innovations, and empirical validation across different market contexts.

7.1 Signal Detection

Signal Detection primarily focuses on extracting robust signals from complex market data. StockFormer [Gao *et al.*, 2023] pioneered a three-branch transformer architecture that captures long-term trends,

Table 9: Summary of Recent Contributions in Portfolio Optimization

Work	Key Innovation	Methodological	Primary Results
MASA [Li *et al.*, 2024e]	Multi-agent framework with specialized risk and return agents	Reward-based guiding mechanism for strategy diversity	Superior risk-adjusted returns across CSI300, DJIA, S&P500
FreQuant [Jeon *et al.*, 2024]	Frequency domain analysis for pattern identification	Multi-granular asset representation through DFT	ARR: 2.1× improvement, Enhanced stability in market shifts
EarnMore [Zhang *et al.*, 2024d]	Customizable portfolio pools with maskable tokens	Self-supervised masking and reconstruction	Profit: +40%, Comparable risk levels
TrendTrader [Ding *et al.*, 2024b]	Multimodal fusion of price and sentiment	Spatial-temporal RL framework	Superior ARR, ASR, MDD in DJIA, SSE-50
Network-EDM [Hui and Wang, 2024]	Extremal risk mitigation through network theory	Maximum independent sets for diversification	Outperformance in market downturns on CSI 300
Market-Graph [Yamagata and Ono, 2024]	Market graph-based clustering for index tracking	Turnover sparsity regularization with primal-dual splitting	Higher Sharpe ratios on S&P500
LLM-Alpha [Kou *et al.*, 2024]	LLM-based alpha factor mining	Multi-agent system with dynamic weight-gating	Return: +53.17% vs -11.73% market (China, 2023)

Table 10: Comparison of Recent Portfolio Optimization Approaches

Model	Adaptive	Risk Management	Training Paradigm	Compute Cost	Market Coverage	Geographic/Asset (Datasets)
MASA	✓	✓	MARL	High	Global	CSI300, DJIA, S&P500
FreQuant	✓	-	RL (DFT)	Medium	Global + Crypto	Global + Crypto
EarnMore	✓	✓	RL + SSL	Low	Multiple Markets	Multiple Markets
TrendTrader	✓	-	RL	High	Regional	DJIA, SSE-50
Network-EDM	-	✓	Optimization	Medium	Regional	China (CSI 300)
Market-Graph	-	✓	Graph Learning	Medium	Regional	US (S&P500)
LLM-Alpha	✓	✓	LLM + RL	Very High	Regional	China (A-shares)

short-term movements, and inter-asset relationships through specialized attention mechanisms. Its diversified multi-head attention design maintains pattern diversity across concurrent time series, demonstrating particular effectiveness in volatile cryptocurrency markets with 40.3% improvement in portfolio returns.

TRR [Zhang *et al.*, 2024c] represents a novel direction in incorporating qualitative data through LLMs, enabling zero-shot reasoning for unprecedented market events. Its four-phase framework combines impact graph generation, temporal context management, and relational reasoning, showing superior performance in detecting market crashes across multiple historical crisis periods.

7.2 Execution Optimization

Execution optimization has evolved toward hierarchical frameworks that decompose complex trading decisions. MacMic [Niu *et al.*, 2024a] introduces a two-level architecture where a high-level agent handles volume scheduling while a low-level agent manages precise order placement. The framework's stacking Hidden Markov Model enables unsupervised extraction of multi-granular market representations, achieving superior price execution across US and Chinese markets.

IMM [Niu *et al.*, 2024b] automates market making through multi-price level strategies, employing temporal-spatial attention networks to mitigate adverse selection risk. Its imitation learning approach from expert strategies facilitates efficient exploration of complex action spaces, demonstrating improved risk-adjusted returns in futures markets.

HRT [Zhao and Welsch, 2024] tackles the joint problem of stock selection and execution through a hierarchical PPO-DDPG framework. Its phased alternating training ensures coordinated learning between controllers, achieving superior performance across both bullish and bearish markets while maintaining portfolio diversification.

7.3 High-Frequency Trading

High-frequency trading systems have advanced through specialized architectures addressing market microstructure and computational efficiency. EarnHFT [Qin *et al.*, 2024] introduces a three-stage framework combining Q-learning with specialized agent pools, effectively handling extended trading trajectories while maintaining

Table 11: Summary of Recent Contributions in Quantitative Trading

Work	Key Innovation	Methodology	Primary Results
StockFormer	Predictive coding with multi-aspect modeling	Three-branch transformer for market dynamics	Returns: +40.3%, SR: +22.7% vs SAC
MacMic	Order execution decomposition	Hierarchical RL with stacking HMM	Superior execution across US/CN markets
IMM	Multi-price market making strategy	Adversarial state representation learning	Lower adverse selection in futures
EarnHFT	Adaptive crypto trading system	Hierarchical Q-learning with router	+30% profit vs SOTA in crypto
MacroHFT	Market regime-based decomposition	Memory-enhanced policy integration	Superior risk-adjusted crypto returns
DRPO	Direct portfolio weight optimization	Probabilistic state decomposition	415ms latency in production
CPPI-MADDPG	Portfolio insurance integration	Multi-agent RL with protection strategies	AR: 9.68%, SR: 2.18 in SZSE
HRT	Integrated selection-execution	PPO-DDPG hierarchy for trading	Improved diversification in S&P500
TRR	LLM-based crash detection	Four-phase temporal reasoning	Superior detection of market crashes

30% higher profitability across market conditions.

MacroHFT [Zong *et al.*, 2024] employs a two-phase approach with specialized sub-agents for different market regimes, using a memory-enhanced hyper-agent for rapid adaptation. DRPO [Han *et al.*, 2023] achieves practical deployment success through state space decomposition and probabilistic dynamic programming, maintaining 415ms latency in production environments.

CPPI-MADDPG [Zhang *et al.*, 2024b] integrates classical portfolio insurance with multi-agent reinforcement learning, achieving 9.68% annual returns while maintaining downside protection through adaptive strategy selection.

7.4 Comparative Analysis

The surveyed approaches demonstrate distinct characteristics in their architectural design, computational requirements, and market applicability. Table 12 summarizes key aspects across different frameworks:

The architectural approaches demonstrate clear trade-offs between modeling sophistication and practical deployment constraints. Predictive frameworks like StockFormer achieve superior performance through comprehensive market modeling but require significant computational resources. In contrast, execution-focused systems like DRPO maintain efficiency through focused optimization while sacrificing broader strategy integration.

Furthermore, compute requirements vary significantly, from DRPO's production-ready 415ms latency to TRR's resource-intensive LLM processing. This latency-sophistication trade-off critically influences deployment scenarios, with HFT systems prioritizing speed while signal generation approaches emphasize modeling capacity.

Market coverage and generalization capabilities also differ substantially. Models like StockFormer and TRR demonstrate cross-market applicability, while specialized frameworks like IMM and EarnHFT show superior performance within specific market contexts. This specialization-generalization trade-off suggests that optimal deployment strategies might involve ensemble approaches tailored to specific market conditions and trading objectives.

8 Knowledge Retrieval and Augmentation

Knowledge retrieval and augmentation in financial markets involves extracting, processing, and leveraging structured and unstructured information for improved decision-making. This section focuses on discussing three primary directions: financial information retrieval pipelines that processes complex financial texts, intelligent report generation that synthesizes multiple data sources, and agent-based market simulation that enables scenario analysis. As summarized in Table 13, these works range from specialized architectures for financial text processing to comprehensive frameworks for market simulation.

8.1 Information Retrieval

Information retrieval (IR) focused on processing complex financial documents through specialized models. MACK [Huang *et al.*, 2024] established a comprehensive framework for Chinese financial event extraction through matrix chunking. Unlike previous approaches requiring pre-identified entities, MACK processes raw text directly through a two-dimensional annotation method that visualizes component interactions. When evaluated on the FINEED dataset with 5,000 annotated events, this architecture achieved 81.33% F1-score in event extraction while maintaining 96.89% accuracy in word segmentation.

The emergence of large language models has enabled new approaches to financial annotation. LLM-Annotator [Aguda *et al.*, 2024] demonstrated that models like GPT-4 and PaLM 2 can outperform crowdworkers by 29% in accuracy while maintaining cost efficiency. Using the REFinD dataset with 28,676 SEC filing relations, the framework's reliability index successfully identified in-

Table 12: Comparison of Quantitative Trading Approaches

Model	Strategy	Execution	Market	Compute	Key Strengths	Key Limitations
StockFormer	Predictive	Indirect	Multi-asset	High	Robust pattern extraction; Volatile market performance	High compute overhead; Complex training
MacMic	Execution	Direct	Equities	Medium	Superior price execution; Hierarchical control	Limited signal generation; Single market focus
IMM	Market Making	Direct	Futures	Medium	Low adverse selection; Queue priority preservation	Market-specific design; Limited asset coverage
EarnHFT	HFT	Direct	Crypto	High	Strong profitability; Adaptive routing	Crypto-specific; Resource intensive
MacroHFT	HFT	Direct	Crypto	High	Regime adaptation; Memory-enhanced decisions	Complex training process; Market-specific
DRPO	HFT	Direct	Multi-asset	Low	Production-ready latency; Efficient computation	Limited strategy scope; Basic signals
CPPI-MADDPG	Portfolio	Indirect	Equities	Medium	Downside protection; Strategy integration	Coordination overhead; Slower adaptation
HRT	Hybrid	Direct	Equities	High	Balanced execution; Portfolio diversification	Complex training; High resource usage
TRR	Signal	Indirect	Multi-asset	Very High	Novel signal sources; Crisis detection	High latency; Resource intensive

Table 13: Summary of Recent Contributions in Knowledge Retrieval and Augmentation

Work	Key Innovation	Methodology	Primary Results
MACK	Matrix chunking for financial event extraction	Two-dimensional annotation method	F1: 81.33% (Event), 96.89% (Word)
FinReport	Automated investment analysis	Three-module system with news factorization	Return: 57.76%, Accuracy: 75.40%
LLM-Annotator	LLM-based financial relation extraction	Reliability index for annotation quality	29% improvement over crowdworkers
TRR	Temporal reasoning for crash detection	Four-phase news analysis framework	Superior crisis detection across multiple periods
StockAgent	Multi-LLM trading simulation	Dynamic agent behavior modeling	Quantified impact of information sources
EconAgent	LLM-powered economic simulation	Perception-memory modules for agents	Reproduction of key economic phenomena

stances requiring expert review, enabling reliable automation of approximately 65% of annotation tasks.

8.2 Report Generation

The synthesis of multiple information sources for automated analysis has advanced through sophisticated architectures. FinReport [Li *et al.*, 2024d] pioneered this direction through a three-module system combining news factorization, return forecasting, and risk assessment. The framework integrates semantic role labeling with dependency parsing for news understanding, while employing enhanced Fama-French models for return prediction. This approach achieved 75.40% accuracy in news factorization and 57.76% annualized returns in backtesting.

TRR [Koa *et al.*, 2024b] advanced the field through temporal relational reasoning for market event detection. The framework's four-phase approach combines impact chain generation, temporal context management, and PageRank-based attention mechanisms. This architecture demonstrated particular effectiveness in detecting market crashes across multiple historical crises, including the 2007 financial crisis and 2020 COVID-19 crash.

8.3 Agentic Simulation

The complexity of market interactions has motivated the development of sophisticated simulation frameworks. StockAgent [Zhang *et al.*, 2024a] established a comprehensive multi-agent system using different LLMs to simulate realistic trading behaviors. The framework incorporates external factors including macroeconomic conditions and market sentiment, revealing distinct trading patterns between models like GPT-3.5 and Gemini Pro. Experimental results quantified the impact of different information sources, with interest rate information significantly affecting trading frequency.

EconAgent [Li *et al.*, 2024a] advanced this direction through LLM-powered agents for macroeconomic simulation. The framework's innovation lies in its perception and memory modules that enable heterogeneous decision-making while considering historical market dynamics. This approach successfully reproduced key eco-

nomic phenomena like the Phillips Curve and demonstrated effectiveness in simulating crisis impacts, particularly during the COVID-19 period.

8.4 Comparative Analysis

Table **??** summarizes key aspects of representative models The three research directions in knowledge retrieval and augmentation offer complementary approaches to market understanding, each with distinct theoretical limitations. Information extraction methods face fundamental challenges in handling context-dependent financial terminology and evolving market jargon. While MACK demonstrates strong performance on Chinese texts, the generalization to multiple languages and domains remains challenging due to linguistic complexity and domain-specific variations.

Report generation frameworks encounter theoretical limitations in causal reasoning and temporal dependency modeling. Although FinReport successfully combines multiple information sources, establishing robust causal relationships between news events and market movements remains an open challenge. The inherent delay between information release and market impact further complicates temporal modeling.

Simulation approaches face fundamental limitations in agent behavior modeling and market complexity. While StockAgent and EconAgent demonstrate impressive capabilities, they necessarily simplify market microstructure and agent interactions. The challenge of calibrating agent behaviors to realistic market conditions while maintaining computational tractability represents a key area for future research.

9 Financial Datasets

Financial datasets play a crucial role in advancing artificial intelligence applications for market analysis and trading. Recent developments in dataset creation have focused on three primary directions: synthetic data generation for enhanced model training, multimodal integration combining market data with textual information, and specialized datasets for reasoning and sentiment analysis. Unlike traditional financial datasets that focus solely on price and volume information, modern datasets incorporate diverse data types including news sentiment, investor emotions, and chain-of-thought reasoning annotations. As summarized in Table 14, these works contribute to both the creation of novel datasets and the development of sophisticated data generation techniques, addressing critical challenges in data availability, quality, and comprehensiveness for financial applications.

9.1 Synthetic Data Generation

Market-GAN [Xia *et al.*, 2024a] established a framework for generating high-fidelity financial data with controllable semantic context through a novel two-stage training approach. The framework combines GAN architecture with autoencoder and supervisory networks to maintain data fidelity while ensuring alignment with given market contexts. Its C-TimesBlock innovation effectively captures temporal dependencies while preventing mode collapse, a common challenge in financial time series generation. When evaluated on Dow Jones Industrial Average data spanning 2000-2023, this approach demonstrated superior performance in context alignment and downstream task fidelity compared to existing generation methods.

9.2 Multimodal Fusion

FNSPID [Dong *et al.*, 2024b] pioneered large-scale integration of quantitative and qualitative financial data through a comprehensive dataset combining 29.7 million stock prices with 15.7 million aligned news records. The dataset covers 4,775 S&P500 companies from 1999-2023, demonstrating that increased dataset scale significantly improves prediction accuracy. Experiments with various deep learning architectures revealed that transformer-based models achieve optimal performance ($R^2 = 0.988$) when leveraging the dataset's full scale, while sentiment information provides modest but consistent improvements in prediction accuracy.

9.3 Sentiment and Emotion

AlphaFin [Li *et al.*, 2024c] addressed the interpretability gap in financial analysis through a dataset combining traditional market data with chain-of-thought annotations. The framework's integration with retrieval-augmented generation enables real-time market awareness while maintaining reasoning capabilities, achieving 30.8% annualized returns in experimental validation. StockEmotions [Lee *et al.*, 2023] complemented this direction through fine-grained emotion analysis, providing 10,000 annotated StockTwits comments across 12 emotion classes. The dataset's multi-step annotation pipeline, combining pre-trained language models with expert validation, demonstrated that incorporating emotional features significantly improves market prediction accuracy compared to purely numerical approaches.

9.4 Comparative Analysis

These financial datasets exhibit distinct characteristics in their coverage, scale, and applicability to different financial tasks. As shown in Table 15, coverage ranges from focused sentiment analysis (StockEmotions) to comprehensive market-wide data integration (FNSPID). Scale varies significantly, with FNSPID providing millions of aligned price-news records while StockEmotions offers deeper but narrower emotional annotation.

Market coverage and temporal resolution reveal complementary strengths across datasets. Market-GAN focuses on high-fidelity synthetic data for a single major index (DJIA) with emphasis on temporal consistency. In contrast, FNSPID provides broader market coverage across the S&P500 with emphasis on news-price alignment. AlphaFin bridges multiple markets while prioritizing reasoning annotation quality, and StockEmotions offers specialized coverage of retail investor sentiment through StockTwits data.

10 Time Series Models

Time series modeling for financial applications focuses on three primary directions: foundation models that leverage large-scale pre-training, efficient architectures that handle temporal complexity, and unified frameworks that address multiple time series tasks. Unlike traditional statistical approaches, modern time series models incorporate transformer architectures, multiscale analysis, and transfer learning capabilities. As summarized in Table 16, these foundation models are trained on massive datasets, specialized architectures for temporal modeling, and unified frameworks supporting multiple tasks.

10.1 Foundation Models

The emergence of foundation models has transformed time series analysis through large-scale pre-training and transfer learning. Timer established a comprehensive framework using a decoder-only transformer architecture trained on 1 billion time points. The model's innovation lies in its unified sequence format that handles heterogeneous time series data while enabling few-shot learning capabilities. When evaluated across forecasting, imputation, and anomaly detection tasks, this architecture achieved state-of-the-art performance using only 1-5MOMENT advanced this direction

Table 14: Summary of Recent Contributions in Financial Datasets

Dataset	Key Innovation	Data Characteristics	Primary Results
Market-GAN	Controllable synthetic data with semantic context	Two-stage GAN with C-TimesBlock for temporal consistency	Superior fidelity in DJIA simulation (2000-2023)
FNSPID	Large-scale price-news integration framework	29.7M prices, 15.7M news records for 4,775 companies	Transformer accuracy R^2 = 0.988, Reproducible updates
AlphaFin	Chain-of-thought financial reasoning	Market data with expert reasoning annotations	30.8% annualized returns, Enhanced interpretability
StockEmotions	Fine-grained investor psychology	10k annotated comments, 12 emotion classes, emoji features	Improved prediction with emotional features

Table 15: Comparison of Financial Dataset Characteristics

Dataset	Time Series	Text Data	Market Coverage	Scale	Primary Application
Market-GAN	✓	-	DJIA	Medium	Market Simulation
FNSPID	✓	✓	S&P500	Very Large	Predictive Modeling
AlphaFin	✓	✓	Multiple	Large	Interpretable Analysis
StockEmotions	✓	✓	StockTwits	Medium	Sentiment Analysis

through an open-source family of foundation models for general-purpose time series analysis. The framework addresses key challenges in multi-dataset training through specialized techniques for handling varying sampling rates and amplitudes. Its contribution of Time Series Pile, a diverse collection of public datasets, enables robust pre-training while maintaining reproducibility. This approach demonstrated particular strength in limited-supervision settings across financial applications.

10.2 Efficient Architectures

The complexity of temporal patterns has motivated development of specialized architectures balancing modeling capacity with computational efficiency. TimeMixer pioneered this direction through a multiscale MLP-based architecture that decomposes temporal variations at different sampling scales. The model employs Past-Decomposable-Mixing blocks for separate processing of seasonal and trend components, while Future-Multipredictor-Mixing blocks combine predictions across scales. This architecture demonstrated state-of-the-art performance across 18 benchmarks while maintaining computational efficiency. PatchTST advanced efficient processing through a patch-based approach to time series modeling. The framework's key innovations lie in subseries-level patching for local information capture and channel-independent processing for univariate sequences. This architecture achieved 21% reduction in Mean Squared Error for long-term forecasting while effectively handling extended historical sequences without memory constraints.

10.3 Unified Frameworks

The diverse requirements of time series analysis have inspired development of unified frameworks supporting multiple tasks. TimesNet established a comprehensive architecture through multi-periodicity analysis that transforms one-dimensional sequences into two-dimensional representations. The framework captures both intraperiod and interperiod variations through parameter-efficient inception blocks, demonstrating superior performance across forecasting, imputation, classification, and anomaly detection tasks.

10.4 Discussion

The three research directions in time series modeling offer complementary approaches to temporal analysis, each with distinct theoretical limitations. Table 17 summarizes key aspects across different frameworks. Foundation models face fundamental challenges in capturing rare temporal patterns and handling non-stationary distributions, though they excel at extracting general temporal dependencies. While large-scale pre-training improves robustness, the inherent challenge of temporal causality and regime changes remains.

Efficient architectures encounter trade-offs between model capacity and computational complexity. While patch-based approaches and multiscale decomposition provide practical solutions, they necessarily introduce approximations in temporal modeling. The balance between local and global temporal dependencies remains a key theoretical challenge.

Unified frameworks face fundamental limitations in optimal parameter sharing across diverse tasks. While architectures like TimesNet demonstrate impressive multi-task capability, the theoretical foundations for optimal architecture design across different temporal modeling objectives remain incomplete. These limitations suggest future research directions in developing more theoretically grounded approaches to temporal modeling while maintaining practical applicability.

The integration of domain knowledge remains crucial for developing more robust and interpretable temporal models, particularly in financial applications where model reliability and theoretical soundness are paramount.

11 Open Challenges

Having comprehensively reviewed the recent works on Financial Ai, we identify several promising research directions that warrant further investigation. In this section, We organize these opportunities into architectural innovations, methodological advancements, and practical considerations to potentially extend the impact to further advancement of the field.

Table 16: Summary of Recent Contributions in Time Series Models

Model	Key Innovation	Methodology	Primary Results
Timer	Pre-trained decoder-only transformer	Unified sequence format for heterogeneous data	Strong few-shot performance with 1-5% data
MOMENT	Open-source foundation model family	Multi-dataset training strategies	Superior limited-supervision performance
TimeMixer	Multiscale MLP architecture	Decomposable mixing blocks for temporal patterns	SOTA across 18 benchmarks
TimesNet	Multi-periodicity analysis	1D to 2D transformation for temporal patterns	Unified performance across 5 major tasks
PatchTST	Patch-based time series processing	Channel-independent transformer	21% MSE reduction in long-term forecasting

Table 17: Comparison of Time Series Model Characteristics

Model	Pre-training	Multi-task	Compute	Scalability	Primary Application
Timer	✓	✓	High	Linear	Few-shot Learning
MOMENT	✓	✓	High	Linear	Limited Supervision
TimeMixer	-	-	Medium	Sublinear	Multiscale Patterns
TimesNet	-	✓	Medium	Linear	Task Flexibility
PatchTST	-	-	Low	Sublinear	Long Sequences

11.1 Architectural Innovations

The evolution of financial AI architectures suggests several key directions for improvement. Foundation models pre-trained on massive financial datasets show promise [Liu *et al.*, 2024b; Goswami *et al.*, 2024] but currently lack domain-specific inductive biases crucial for finance. Future research should explore specialized pre-training objectives that incorporate market microstructure theory and regulatory constraints. The development of modular architectures that combine pre-trained components with task-specific adaptors could enable more efficient transfer learning while maintaining model interpretability. Multi-agent architectures for portfolio optimization and market making demonstrate strong performance [Li *et al.*, 2024e; Zhang *et al.*, 2024b] but face challenges in convergence guarantees and Nash equilibrium stability. Research into theoretical frameworks for multi-agent learning under market non-stationarity could lead to more robust trading systems. The integration of market impact models into agent training frameworks, building upon work like [Niu *et al.*, 2024a], remains crucial for bridging the gap between simulation and real-world deployment.

11.2 Methodological Advancements

Several methodological challenges require attention from the research community. The treatment of temporal dependencies in financial data remains suboptimal, with current approaches [Wang *et al.*, 2024a; Nie *et al.*, 2023] struggling to capture long-range dependencies while maintaining computational efficiency. Future work should explore adaptive attention mechanisms that dynamically adjust their receptive field based on market conditions. The integration of classical financial theory with deep learning frameworks presents another promising direction. While works like [Xu *et al.*, 2024] incorporate GARCH models and factor analysis, a more systematic approach to theory-guided neural network design could improve both performance and interpretability. This includes developing loss functions that explicitly encode financial principles and constraints. Risk modeling in deep learning frameworks requires particular attention. Current approaches [Hui and Wang, 2024; Yamagata and Ono, 2024] often rely on simple volatility estimates or Sharpe ratios, failing to capture tail risks and regime changes adequately. Research into neural architectures specifically designed for extreme value theory and systemic risk modeling could significantly improve portfolio management and risk assessment.

11.3 Multimodal Foundation Models

The emergence of multimodal foundation models presents exciting opportunities for financial applications. While current work [Li *et al.*, 2024c; Zhang *et al.*, 2024a] demonstrates promise in combining textual and numerical data, future research should explore more sophisticated fusion mechanisms. This includes developing specialized architectures for processing earnings calls, satellite imagery, and alternative data sources simultaneously. The adaptation of foundation models for real-time market analysis remains challenging. Research into streaming architectures that can efficiently update their knowledge base and adapt to new market conditions could enable more responsive trading systems. Additionally, exploring techniques for maintaining temporal coherence across different data modalities could improve prediction accuracy during market regime changes.

11.4 Hardware-Accelerated Trading

The advancement of hardware-accelerated trading systems presents several research opportunities. Current approaches [Han *et al.*, 2023; Qin *et al.*, 2024] demonstrate promising potential for low-latency trading. Research into specialized neural network architectures optimized for FPGA implementation could further reduce latency while maintaining model sophistication. The development of hardware-aware training algorithms represents another promising direction. Future work should explore techniques for jointly optimizing model architecture and hardware implementation, potentially leading to more efficient high-frequency trading systems. This includes developing specialized attention mechanisms and network

pruning techniques that consider hardware constraints during training. The integration of hardware acceleration with market making systems presents unique challenges. Research into architectures that can efficiently process order book data and execute trades with microsecond latency could significantly improve market liquidity. This includes developing specialized circuits for order matching and risk calculation that maintain accuracy while minimizing latency.

11.5 Research-Industry Collaborations

A notable limitation in current research is the absence of real-world deployment studies and industry validation. While works like [Han *et al.*, 2023] report production latency metrics, and [Qin *et al.*, 2024; Zong *et al.*, 2024] demonstrate strong backtesting performance, none of the surveyed papers provide comprehensive evidence of successful industrial deployment. This gap manifests in several critical aspects: First, most research evaluations rely solely on historical data backtesting, which fails to capture crucial real-world challenges such as market impact, execution slippage, and microstructure effects. The lack of live trading results or paper trading validation raises questions about the practical viability of proposed methods. Second, computational requirements and system integration receive limited attention. While papers like [Han *et al.*, 2023] discuss latency constraints, few address critical production concerns such as system reliability, fault tolerance, and integration with existing trading infrastructure. The absence of deployment case studies or performance analysis under real market conditions represents a significant limitation in current research. Third, regulatory compliance and risk management frameworks remain largely theoretical. Despite works like [Li *et al.*, 2024d] addressing interpretability, none of the surveyed papers demonstrate compliance with specific regulatory requirements or integration with institutional risk management systems. This gap between academic innovation and regulatory reality hinders industrial adoption. Future research would benefit significantly from collaborations between academic institutions and financial firms to validate proposed methods under real market conditions. Studies documenting deployment challenges, system architecture decisions, and practical performance metrics would provide valuable insights for both researchers and practitioners. Additionally, research into robust evaluation frameworks that better approximate real-world trading conditions could help bridge the gap between academic research and industrial applications.

11.6 Practical Considerations

Several critical implementation challenges require focused research attention. While synthetic data generation techniques [Xia *et al.*, 2024a] help address data limitations, research is needed on frameworks that can rigorously validate model robustness under real market conditions. This includes developing standardized stress testing methodologies that simulate market crises, liquidity shocks, and extreme events. System resilience and failsafe mechanisms represent another crucial area. Research into graceful degradation strategies, automated circuit breakers, and robust fallback mechanisms could improve the safety of automated trading systems. This includes developing methods to automatically detect and respond to anomalous market conditions or model behavior. The challenge of continuous learning and model updating in production environments remains largely unexplored. Future research should investigate techniques for safely updating models in live trading systems, including methods for gradual deployment, A/B testing, and performance monitoring. This includes developing frameworks for detecting model degradation and safely reverting changes if necessary. Cross-jurisdictional compliance presents another significant challenge, particularly for global trading operations. Research into automated compliance verification and real-time regulatory reporting could help bridge the gap between academic innovation and practical deployment. This includes developing standardized interfaces between trading systems and regulatory monitoring frameworks. These practical considerations highlight the need for research that explicitly addresses production deployment challenges while maintaining the theoretical rigor expected in academic work. Success in these areas could accelerate the adoption of AI innovations in real-world financial systems.

12 Conclusion

This survey has systematically analyzed recent advancements in artificial intelligence for financial applications, revealing significant progress across predictive modeling, decision-making, and knowledge retrieval tasks. The examined innovations span architectural developments in foundation models, specialized network designs for temporal and relational learning, and practical frameworks for production deployment. While these advances demonstrate substantial improvements in model performance and capability, several crucial challenges remain unresolved.

The theoretical foundations for financial AI systems still require strengthening, particularly in establishing convergence guarantees for multi-agent trading systems and developing robust frameworks for handling market non-stationarity. The integration of classical financial theory with deep learning architectures presents another critical area for development, as current approaches often fail to fully incorporate market microstructure effects and regulatory constraints.

Production deployment remains a significant challenge, with few studies demonstrating successful real-world implementation or comprehensive compliance with regulatory requirements. The gap between academic innovation and industrial application suggests the need for closer collaboration between researchers and practitioners, particularly in developing standardized evaluation frameworks that better approximate real-world trading conditions.

Future research directions should focus on developing more robust theoretical frameworks for financial AI, improving the integration of domain knowledge in model architecture design, and addressing the practical challenges of system deployment and regulatory compliance. The advancement of hardware-accelerated trading systems and the development of multimodal foundation models present particularly promising opportunities for innovation. Success in these areas could significantly advance the field while improving the reliability and effectiveness of AI-driven financial systems.

References

[Aguda *et al.*, 2024] Toyin Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, Charese Smiley, and Sameena Shah. Large language models as financial data annotators: A study on effectiveness and efficiency. *arXiv preprint arXiv:2403.18152*, 2024.

[Chen *et al.*, 2024] Weijun Chen, Shun Li, Xipu Yu, Heyuan Wang, Wei Chen, and Tengjiao Wang. Automatic de-biased temporal-relational modeling for stock investment recommendation. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 1999–2008. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Daiya *et al.*, 2024] Divyanshu Daiya, Monika Yadav, and Harshit Singh Rao. Diffstock: Probabilistic relational stock market predictions using diffusion models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7335–7339. IEEE, 2024.

[Ding *et al.*, 2024a] Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*, 2024.

[Ding *et al.*, 2024b] Wei Ding, Zhennan Chen, Hanpeng Jiang, Yuanguo Lin, and Fan Lin. Trend-heuristic reinforcement learning framework for news-oriented stock portfolio management. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5120–5124. IEEE, 2024.

[Dong *et al.*, 2024a] Michael Ming Dong, Theophanis C Stratopoulos, and Victor Xiaoqi Wang. A scoping review of chatgpt research in accounting and finance. *SSRN*, 2024.

[Dong *et al.*, 2024b] Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4918–4927, 2024.

[Gao *et al.*, 2023] Siyu Gao, Yunbo Wang, and Xiaokang Yang. Stockformer: Learning hybrid trading machines with predictive coding. In *IJCAI*, pages 4766–4774, 2023.

[Goswami *et al.*, 2024] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

[Han *et al.*, 2023] Li Han, Nan Ding, Guoxuan Wang, Dawei Cheng, and Yuqi Liang. Efficient continuous space policy optimization for high-frequency trading. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4112–4122, 2023.

[Huang *et al.*, 2024] Yusheng Huang, Ning Hu, Kunping Li, Nan Wang, and Zhouhan Lin. Extracting financial events from raw texts via matrix chunking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7035–7044, 2024.

[Hui and Wang, 2024] Qian Hui and Tiandong Wang. Mitigating extremal risks: A network-based portfolio strategy. *arXiv preprint arXiv:2409.12208*, 2024.

[Jain *et al.*, 2024] Samyak Jain, Parth Chhabra, Atula Tejaswi Neerkaje, Puneet Mathur, Ramit Sawhney, Shivam Agarwal, Preslav Nakov, Sudheer Chava, and Dinesh Manocha. Saliency-aware interpolative augmentation for multimodal financial prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14285–14297, 2024.

[Jeon *et al.*, 2024] Jihyeong Jeon, Jiwon Park, Chanhee Park, and U Kang. Frequant: A reinforcement-learning based adaptive portfolio optimization with multi-frequency decomposition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1211–1221, 2024.

[Koa *et al.*, 2024a] Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4304–4315, 2024.

[Koa *et al.*, 2024b] Kelvin JL Koa, Yunshan Ma, Ritchie Ng, Huanhuan Zheng, and Tat-Seng Chua. Temporal relational reasoning of large language models for detecting stock portfolio crashes. *arXiv preprint arXiv:2410.17266*, 2024.

[Kou *et al.*, 2024] Zhizhuo Kou, Holam Yu, Jingshu Peng, and Lei Chen. Automate strategy finding with llm in quant investment. *arXiv preprint arXiv:2409.06289*, 2024.

[Lee *et al.*, 2023] Jean Lee, Hoyoul Luis Youn, Josiah Poon, and Soyeon Caren Han. Stockemotions: Discover investor emotions for financial sentiment analysis and multivariate time series. *arXiv preprint arXiv:2301.09279*, 2023.

[Lee *et al.*, 2024] Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*, 2024.

[Li *et al.*, 2023] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382, 2023.

[Li *et al.*, 2024a] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, 2024.

[Li *et al.*, 2024b] Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. Master: Market-guided stock transformer for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 162–170, 2024.

[Li *et al.*, 2024c] Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, Jun Huang, and Wei Lin. Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. *arXiv preprint arXiv:2403.12582*, 2024.

[Li *et al.*, 2024d] Xiangyu Li, Xinjie Shen, Yawen Zeng, Xiaofen Xing, and Jin Xu. Finreport: Explainable stock earnings forecasting via news factor analyzing model. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 319–327, 2024.

[Li *et al.*, 2024e] Zhenglong Li, Vincent Tam, and Kwan L. Yeung. Developing a multi-agent and self-adaptive framework with deep reinforcement learning for dynamic portfolio risk management. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 1174–1182, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems.

[Liu *et al.*, 2024a] Mengpu Liu, Mengying Zhu, Xiuyuan Wang, Guofang Ma, Jianwei Yin, and Xiaolin Zheng. Echo-gl: Earnings calls-driven heterogeneous graph learning for stock movement prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13972–13980, 2024.

[Liu *et al.*, 2024b] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Forty-first International Conference on Machine Learning*, 2024.

[Nie *et al.*, 2023] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[Niu *et al.*, 2024a] Hui Niu, Siyuan Li, and Jian Li. Macmic: Executing iceberg orders via hierarchical reinforcement learning. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6008–6016. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Niu *et al.*, 2024b] Hui Niu, Siyuan Li, Jiahao Zheng, Zhouchi Lin, Bo An, Jian Li, and Jian Guo. Imm: An imitative reinforcement learning approach with predictive representation learning for automatic market making. In Kate Larson, editor, *Proceedings of*

the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pages 5999–6007. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Qin *et al.*, 2024] Molei Qin, Shuo Sun, Wentao Zhang, Haochong Xia, Xinrun Wang, and Bo An. Earnhft: Efficient hierarchical reinforcement learning for high frequency trading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14669–14676, 2024.

[Wang and Ma, 2024] Mengyu Wang and Tiejun Ma. Mananet: Mitigating aggregated sentiment homogenization with news weighting for enhanced market prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2379–2389, 2024.

[Wang *et al.*, 2024a] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

[Wang *et al.*, 2024b] Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 2024.

[Xia *et al.*, 2024a] Haochong Xia, Shuo Sun, Xinrun Wang, and Bo An. Market-gan: Adding control to financial market data generation with semantic context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15996–16004, 2024.

[Xia *et al.*, 2024b] Hongjie Xia, Huijie Ao, Long Li, Yu Liu, Sen Liu, Guangnan Ye, and Hongfeng Chai. Ci-sthpan: Pre-trained attention network for stock selection with channel-independent spatio-temporal hypergraph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9187–9195, 2024.

[Xiang *et al.*, 2024] Quanzhou Xiang, Zhan Chen, Qi Sun, and Rujun Jiang. Rsap-dfm: Regime-shifting adaptive posterior dynamic factor model for stock returns prediction. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6116–6124. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Xu *et al.*, 2024] Zeda Xu, John Liechty, Sebastian Benthall, Nicholas Skar-Gislinge, and Christopher McComb. Garch-informed neural networks for volatility prediction in financial markets. *arXiv preprint arXiv:2410.00288*, 2024.

[Yamagata and Ono, 2024] Eisuke Yamagata and Shunsuke Ono. Risk-managed sparse index tracking via market graph clustering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9796–9800. IEEE, 2024.

[Yan and Tan, 2024] Wenbo Yan and Ying Tan. Double-path adaptive-correlation spatial-temporal inverted transformer for stock time series forecasting. *arXiv preprint arXiv:2409.15662*, 2024.

[You *et al.*, 2024] Zinuo You, Pengju Zhang, Jin Zheng, and John Cartlidge. Multi-relational graph diffusion neural network with parallel retention for stock trends classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6545–6549. IEEE, 2024.

[Zeng *et al.*, 2024] Liang Zeng, Lei Wang, Hui Niu, Ruchen Zhang, Ling Wang, and Jian Li. Trade when opportunity comes: Price movement forecasting via locality-aware attention and iterative refinement labeling. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6134–6142. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Zhang *et al.*, 2024a] Chong Zhang, Xinyi Liu, Mingyu Jin, Zhongmou Zhang, Lingyao Li, Zhengting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. *arXiv preprint arXiv:2407.18957*, 2024.

[Zhang *et al.*, 2024b] Hengxi Zhang, Zhendong Shi, Yuanquan Hu, Wenbo Ding, Ercan E Kuruoğlu, and Xiao-Ping Zhang. Optimizing trading strategies in quantitative markets using multi-agent reinforcement learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 136–140. IEEE, 2024.

[Zhang *et al.*, 2024c] Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4314–4325, 2024.

[Zhang *et al.*, 2024d] Wentao Zhang, Yilei Zhao, Shuo Sun, Jie Ying, Yonggang Xie, Zitao Song, Xinrun Wang, and Bo An. Reinforcement learning with maskable stock representation for portfolio management in customizable stock pools. In *Proceedings of the ACM on Web Conference 2024*, pages 187–198, 2024.

[Zhao and Welsch, 2024] Zijie Zhao and Roy E Welsch. Hierarchical reinforced trader (hrt): A bi-level approach for optimizing stock selection and execution. *arXiv preprint arXiv:2410.14927*, 2024.

[Zhao *et al.*, 2022] Yu Zhao, Huaming Du, Ying Liu, Shaopeng Wei, Xingyan Chen, Fuzhen Zhuang, Qing Li, and Gang Kou. Stock movement prediction based on bi-typed hybrid-relational market knowledge graph via dual attention networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8559–8571, 2022.

[Zhao *et al.*, 2023] Lifan Zhao, Shuming Kong, and Yanyan Shen. Doubleadapt: A meta-learning approach to incremental learning for stock trend forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3492–3503, 2023.

[Zhao *et al.*, 2024] Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*, 2024.

[Zheng *et al.*, 2023] Zetao Zheng, Jie Shao, Jia Zhu, and Heng Tao Shen. Relational temporal graph convolutional networks for ranking-based stock prediction. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 123–136, 2023.

[Zong *et al.*, 2024] Chuqiao Zong, Chaojie Wang, Molei Qin, Lei Feng, Xinrun Wang, and Bo An. Macrohft: Memory augmented context-aware reinforcement learning on high frequency trading. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4712–4721, 2024.