Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Position Article:

Large investment model

Jian GUO‡, Heung-Yeung SHUM

IDEA Research, International Digital Economy Academy, Shenzhen 518045, China

E-mail: guojian@idea.edu.cn; hshum@idea.edu.cn

Received Apr. 27, 2025; Revision accepted Aug. 19, 2025; Crosschecked Oct. 10, 2025

Abstract: Traditional quantitative investment research is encountering diminishing returns alongside rising labor and time costs. To overcome these challenges, we introduce the large investment model (LIM), a novel research paradigm designed to enhance both performance and efficiency at scale. LIM employs end-to-end learning and universal modeling to create an upstream foundation model, which is capable of autonomously learning comprehensive signal patterns from diverse financial data spanning multiple exchanges, instruments, and frequencies. These "global patterns" are subsequently transferred to downstream strategy modeling, optimizing performance for specific tasks. We detail the system architecture design of LIM, address the technical challenges inherent in this approach, and outline potential directions for future research.

Key words: Artificial general intelligence; End-to-end; Large investment model; Quantitative investment; Foundation model; Multimodal large language model

https://doi.org/10.1631/FITEE.2500268 **CLC number:** TP391; F830.59

1 Introduction

Quantitative investment (quant) involves financial investment strategies driven by mathematical, statistical, or machine learning models, and it uses powerful computers to execute trading instructions derived from quant models at speeds and frequencies unattainable by human traders. In particular, deep learning techniques are widely applied in quant modeling, such as stock/futures trend prediction (Zhang L et al., 2017; Xu and Cohen, 2018; Hu Z et al., 2018; Feng et al., 2021), stock selection (Feng et al., 2019; Sawhney et al., 2020, 2021), portfolio optimization (Wang J et al., 2019; Wang Z et al., 2021; Zhang Y et al., 2022; Liu et al., 2023), and algorithmic trading (Fang et al., 2021, 2023; Lin and Beling, 2021; Sun S et al., 2022; Qin et al., 2024).

The traditional quantitative research paradigm is fraught with several limitations. First, it adheres to a comprehensive pipeline that includes data processing, factor mining, machine learning, portfolio optimization, and algorithmic trading. Each of these steps demands significant research resources, including intensive labor and substantial time, to identify effective "alphas." Furthermore, the optimization objectives across these pipeline stages often lack consistency, leading to suboptimal outcomes for the final trading strategy. Additionally, traditional task-specific quantitative modeling relies heavily on pre-defined scenarios, strategy tasks, and associated data, making it difficult to transfer these models smoothly to other strategy tasks. This reliance on "local" data not only limits the model's potential but also exacerbates research costs, as quants are compelled to develop a distinct model for each strategy.

In recent years, the rapid advancements in artificial general intelligence (AGI) have provided a unique opportunity to transform the quantitative research paradigm. Specifically, we discuss the shift toward a new modeling paradigm aimed at enhancing the efficiency and effectiveness of quantitative

‡ Corresponding author
ORCID: Jian GUO, https://orcid.org/0009-0003-5046-2588; Heung-Yeung SHUM, https://orcid.org/0000-0002-4684-911X

finance research. First, there is a clear transition from traditional multifactor modeling to state-of-the-art end-to-end modeling. Unlike multifactor modeling, which builds trading strategies incrementally through a research pipeline, end-to-end modeling seeks to directly generate the final trading strategy, bypassing intermediate steps such as factor mining, and producing predicted alphas, optimal positions, or even algorithmic trading orders. This approach has the potential to eliminate the labor-intensive factor mining process and significantly enhance the efficiency of quantitative research. Second, the shift from traditional task-specific modeling to universal modeling, akin to the "pretrained foundation model + fine-tuned task model" approach commonly used in large language models (LLMs), is becoming increasingly prominent in quantitative investment. The foundation model, typically a universal model trained on a broad and diverse dataset (e.g., data spanning various countries, security markets, and trading assets), can be fine-tuned to optimize specific trading strategies. By combining the strengths of end-to-end modeling and universal modeling, we propose the large investment model (LIM), a novel methodological framework for quantitative investment research. Fig. 1 illustrates the distinctions among multifactor modeling, end-to-end modeling, and universal modeling.

The remainder of this article is organized as Fig. 2. Section 2 provides a brief review of the data, strategies, and research pipeline in quantitative investment. Section 3 introduces the general framework of LIM and illustrates its practical applicability. Details of the upstream foundation modeling and downstream strategy modeling within LIM are presented in Sections 4 and 5, respectively. Section 6 discusses the architecture design for automated strategy generation and trading using LIM. Section 7 proposes several new research directions, and Section 8 offers concluding remarks.

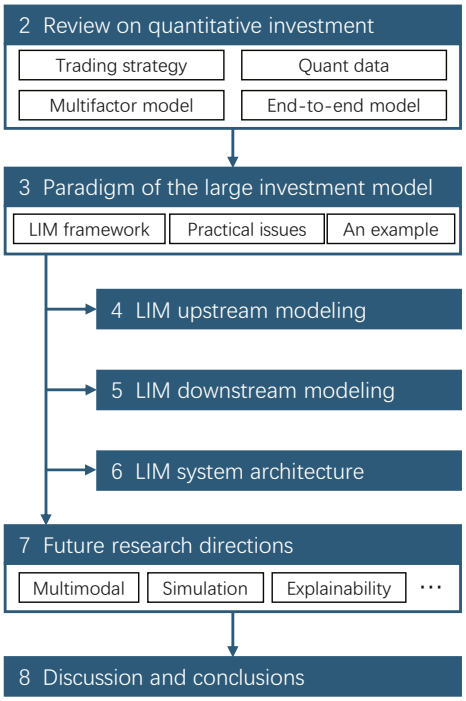

Fig. 2 Section organization for introducing the large investment model (LIM)

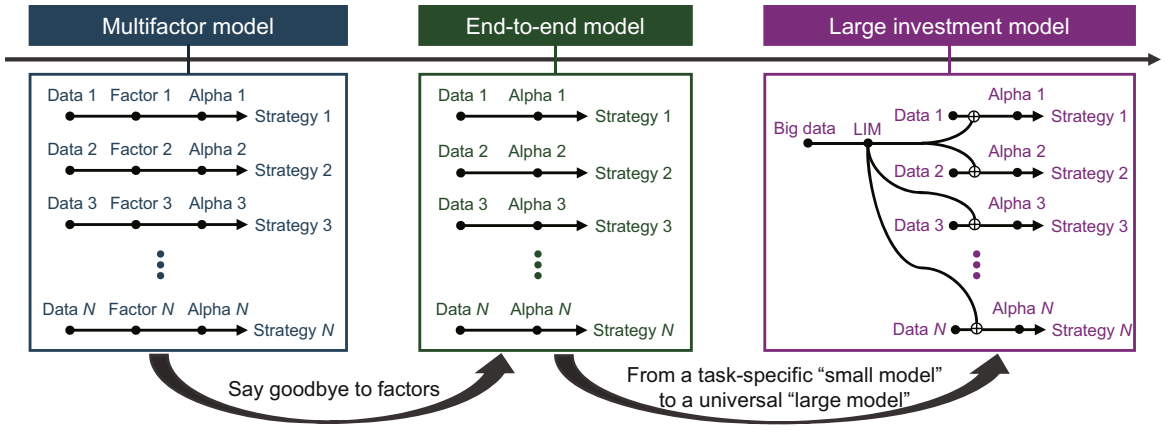

Fig. 1 Three quant research paradigms: multifactor model, end-to-end model, and large investment model (LIM)

2 Review on quantitative investment

Quantitative investment relies on automated strategies built on various data to trade different instruments such as stocks (Bodie et al., 2014), futures (Hull, 2021), bonds (Fabozzi et al., 2021), exchange-traded funds (ETFs) (Gastineau, 2001), and options (Black and Scholes, 1973). This section briefly introduces common concepts, strategies, data, and modeling paradigms in quantitative investment.

2.1 Quant strategies

A quantitative strategy is a systematic function or trading methodology used for trading financial instruments in financial markets. This strategy is based on either predefined rules or trained models for making trading decisions, and is typically the core intellectual property of a trading firm. A standard quantitative strategy should specify several configurations, such as the universe of financial instruments to be traded (the pool of assets screened for a quant strategy, filtered by investability criteria) (Chincarini and Kim, 2022), the average holding period (the duration over which a security remains in a portfolio before being sold), and the trading frequency (how often trades are executed). Additionally, it should define the types of strategies employed. Fig. 3 presents a strategy matrix that illustrates many popular trading strategy examples. The horizontal axis introduces a variety of financial instruments, including stocks, ETFs, futures, options, bonds, foreign exchange (forex), and cryptocurrencies (crypto). The vertical axis contains four types of common trading approaches, each representing standard operations for trading various financial instruments and forming different strategies.

Directional trading is a family of strategies used in financial markets that involve taking a position based on the anticipated direction of a security's price movement and profiting from these price changes by buying or selling securities accordingly. Popular directional trading strategies include trend-following trading (identifying and following the direction of an existing trend, taking long positions in uptrends and short positions in downtrends) (Kim et al., 2016), breakout trading (taking positions when the price decisively moves beyond support or resistance levels, with long positions on bullish breakouts and short positions on bearish breakouts) (Păuna, 2019), and contrarian trading (trading against the market trend by identifying overbought or oversold conditions and taking reverse positions) (Chan, 1988).

Long–short trading (Kwan, 1999), commonly used in hedge funds, is a set of strategies that involve taking both long and short positions in different securities to counteract market volatility effects (the "beta" return) from the overall return and to profit from the "alpha" return. A popular example is stock long–short selection, which predicts the "best" stocks to buy (or long) and the "worst" stocks to sell (or short) at each cross-section over time.

Arbitrage trading (Shleifer and Vishny, 1997) is a family of strategies that exploit price discrepancies between different markets or financial instruments to achieve risk-free profits. Common arbitrage strategies include cross-exchange arbitrage (profiting from price discrepancies of a security across various exchanges), triangle arbitrage (typically used in forex and cryptocurrency markets to exploit exchange rate

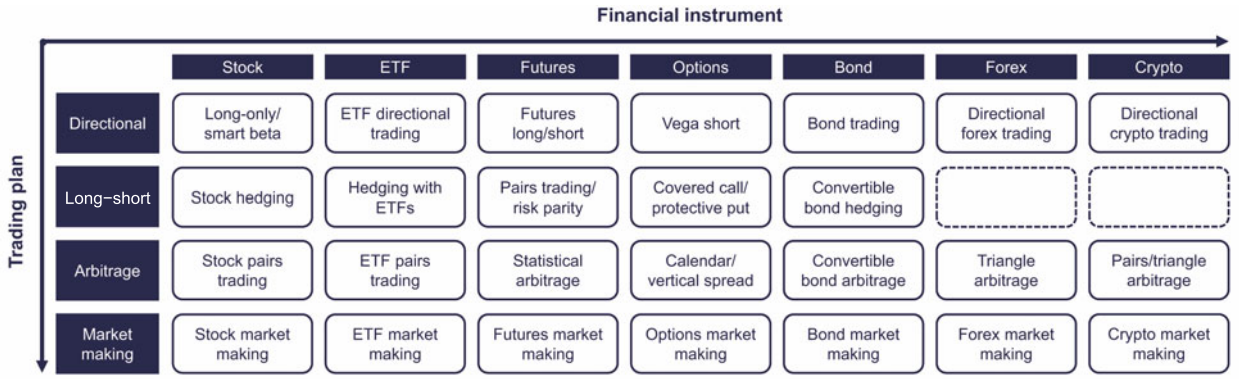

Fig. 3 A strategy matrix for quantitative investment

discrepancies by trading three different currencies), calendar spread arbitrage (for futures), convertible arbitrage (taking a long position in a convertible bond while shorting its underlying stock), and statistical arbitrage (trading pairs of correlated securities by longing the underpriced one and shorting the overpriced one).

Market making trading (Ho and Stoll, 1981) is a family of high-frequency strategies that provide liquidity to financial markets by continuously quoting both buy (bid) and sell (ask) prices for financial instruments, aiming to profit from the spread between these prices. For instance, a market maker might quote a bid price of \$100 and an ask price of \$100.10 for a stock. The market maker buys shares from a trader willing to sell at \$100 and sells them to another trader willing to buy at \$100.10, thereby profiting from the spread \$0.10.

Additionally, trading frequency defines the duration for which assets are held in a portfolio and how often trades are executed. High-frequency trading typically involves holding positions for a few minutes or seconds, whereas low-frequency trading may involve holding assets for several months or years. The significant difference in holding periods between high-frequency and low-frequency trading leads to distinct considerations in strategy design. For example, asset capacity limitations and trading costs are critical issues in high-frequency trading, while managing drawdown risk is a primary concern in low-frequency trading.

2.2 Data diversity in quant modeling

Modern quantitative investment harnesses a diverse array of data to develop statistical and machine learning strategies which aim at profitable trading. Fig. 4 categorizes various types of financial data along two orthogonal dimensions: data depth and data breadth. Data depth refers to the granularity of data, which can span from several years at the macro level to mere nanoseconds at the micro level. Data breadth, on the other hand, indicates the diversity of the data and encompasses the following: quote data such as price/volume, limit order book (LOB) (Madhavan, 2000), and market order flow (Toth et al., 2012); fundamental data such as financial statements (Penman, 2013), investment research reports (Cho et al., 2021), company announcements (MacKinlay, 1997), and analysts' opinions (Chen Q et al., 2005). Data breadth also includes a broad range of alternative data (Sun Y et al., 2024) such as e-commerce transactions, credit card/e-payment transactions (Gupta et al., 2022), news and social media comments (Tetlock et al., 2008), satellite imagery (Yu et al., 2023), foot traffic (Liew et al., 2020), and supply chain data (Paatela et al., 2017).

Different investment strategies rely on different types of financial data. For instance, high-frequency market-making strategies (Ait-Sahalia and Saglam, 2024) focus on data depth by modeling granular LOB data to predict price movements over very short horizons. Horizontal spread arbitrage strategies (Kou et al., 2013) use high-frequency price/volume data to capitalize on price discrepancies in derivatives contracts (such as options or futures) with varying expiration dates. Stock technical trading strategies (Lo et al., 2000) employ candlestick data (open, high, low, and close prices) and trading volumes to generate buy/sell signals. Meanwhile, stock fundamental investing strategies (Wafi et al., 2015) analyze financial statements, analyst reports, and news data to evaluate the fundamental health and intrinsic value of public companies. The rapid growth of Internet and mobile technologies over the past decade has led to an explosion in the accumulation of big data. Financial institutions are increasingly integrating alternative data, such as credit card transaction data, web traffic data, and geolocation data, into their fundamental analysis and value investing strategies (Cao et al., 2024; Dessaint et al., 2024). This shift has significantly expanded the data breadth available for quantitative investment, enabling more comprehensive and nuanced analyses.

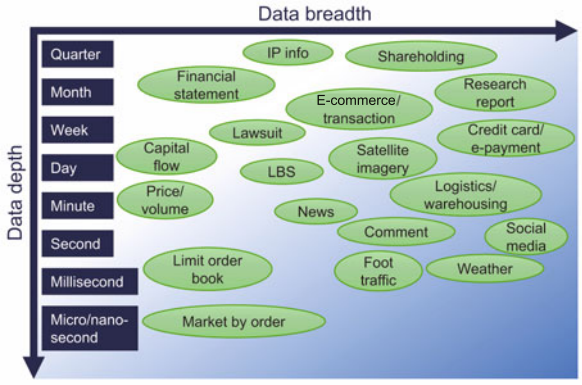

Fig. 4 Spectrum of various financial data along depth and breadth

2.3 Multifactor quant modeling

The quantitative research pipeline comprises several critical stages, including data processing, factor mining, alpha modeling, portfolio position optimization, and order execution optimization, as depicted in the blue section of Fig. 5. Among these stages, factor mining is particularly vital, because the quality of the factors significantly influences the performance of the alpha prediction model, which subsequently impacts the overall returns of the final portfolio.

Factors are typically mathematical formulas or functions that capture signals predictive of trends in various financial instruments, such as stocks, futures, and foreign exchange (Dixon M, 2022). These factors can be derived from a wide range of data sources, including financial quote data, fundamental data, and alternative data. Traditionally, trading factors have been manually designed and constructed, relying heavily on market observations and the expertise of traders. However, there has been a growing shift toward the use of automatic factor mining techniques, such as genetic programming (Chen T et al., 2021), reinforcement learning (Zhao et al., 2025), and LLMs (Wang S et al., 2025), to improve the efficiency of factor construction and selection. Whether manually crafted or algorithmically discovered, these factors undergo rigorous backtesting (a method to evaluate a quantitative strategy's performance by simulating its execution on historical market data under assumed transaction rules and market conditions) (Lo, 2010). Only those factors that are both effective ("good") and non-correlated ("diverse") are retained and stored in the database for use in alpha modeling.

2.4 End-to-end modeling for LIM

Because factors are essentially "features" that characterize instruments, their information is entirely derived from the original data. A natural question arises: can we build a predictive model without explicitly creating factors? The advent of deep learning and end-to-end training paradigm presents a plausible technical route. End-to-end training refers to a modeling approach that directly learns the complex function that links raw inputs to final outputs and encompasses all intermediate stages. Fig. 5 illustrates three types of end-to-end modeling, each starting from the original meta-data (raw data with standardized and simple preprocessing) and leading to different outputs: alpha predictions (e.g., predicted returns over a future horizon), portfolio positions (e.g., the optimal position size at the next trading point), or trade orders (e.g., the optimal order size in the next second for trading).

Recent literature has seen a growing interest in this area, with notable contributions such as deep inception networks (DINs) (Liu et al., 2023) and end-to-end active investment (E2EAI) (Wei et al., 2023). DINs introduce a fully data-driven approach, extracting both time-series and cross-sectional features directly from daily price returns, thereby eliminating

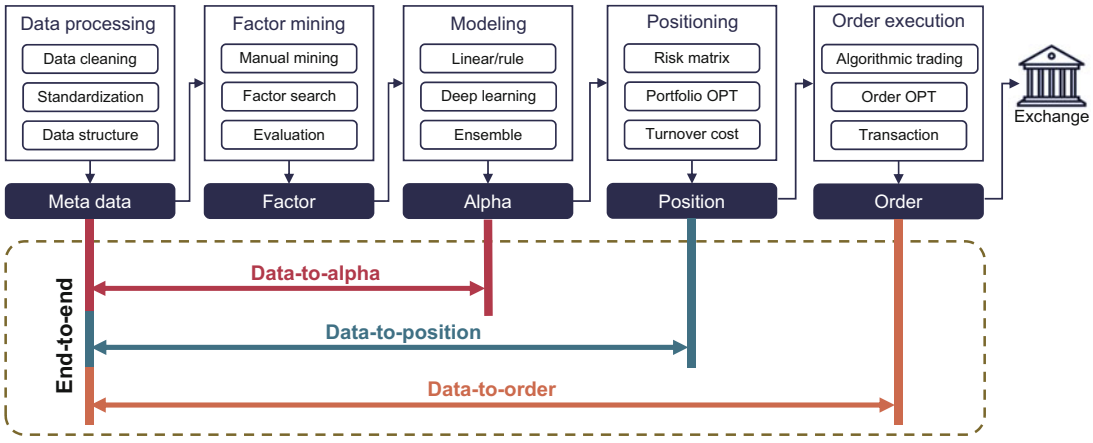

Fig. 5 Quant research pipeline and various types of end-to-end quant modeling (References to color refer to the online version of this figure)

the need for handcrafted features and entirely removing the manual factor mining step. The models are trained to optimize the Sharpe ratio of the entire portfolio, with additional loss terms to balance turnover and systemic risk, and demonstrate strong performance and robustness across multiple asset classes. E2EAI similarly employs a neural network that spans the entire quantitative research pipeline, from factor selection to portfolio construction, with all stages learned jointly via a portfolio-level objective, but it typically starts from a pre-defined factor or meta-factor database as input. Both approaches bypass traditional workflows by directly outputting optimal positions, yet only DIN operates end to end from raw prices without reliance on pre-constructed factors. Another significant contribution is DeepLOB (Zhang Z et al., 2019), which constructs a large-scale deep learning model to predict price movements directly from LOB data of cash equities. Notably, the authors emphasize that DeepLOB generalizes well to instruments not included in the training set, demonstrating the model's ability to extract universal features. The team further extends deep learning models to more granular micro-structure data (Zhang Z et al., 2021), concluding that an ensemble of market by order data and LOB data enhances forecasting accuracy. Unlike the straightforward end-to-end modeling in Zhang Z et al. (2019, 2021), Jiao et al. (2023) proposed an upstream pretraining framework to extract alphas from order flow data, applicable across various granularities and scenarios. This approach also inspires the LIM proposed in this paper.

Compared with the traditional quant research pipeline (blue part of Fig. 5), end-to-end modeling offers several advantages: (1) In traditional quantitative research, the optimization goals of individual modules are usually inconsistent. For instance, each factor is evaluated and selected based on criteria that primarily concern the factor itself, rather than its interaction with other factors. As a result, a "good" factor with a high information coefficient (IC) (Zhang F et al., 2020) or Sharpe ratio (Sharpe, 1966) may negatively impact an alpha model due to complex interactions with other factors, whereas a "bad" factor might significantly contribute to the model. (2) The formulaic nature and operator space of factors can limit their representational capability. Almost all operators (e.g., rank(·) and ts_max(·) as described in Kakushadze (2016)) that define formulaic factors are simple algebraic functions, and the representational power of their combinations is difficult to compare with deep neural networks. Therefore, with sufficient sample size, end-to-end modeling has a higher ceiling than traditional multifactor modeling. (3) Factor mining is a labor-intensive and time-consuming process, especially for building and selecting factors by hand. On the contrary, end-to-end modeling throws such "dirty work" to deep learning algorithms and may reduce the cost significantly.

3 Large investment model

The success of LLMs (Brown T et al., 2020; OpenAI et al., 2024) has highlighted the extraordinary potential of self-supervised generative learning (Bengio et al., 2013), which is grounded in a universal "next-token prediction" task (sequence-to-sequence prediction). In this paradigm, a general-purpose foundation model (Zhou et al., 2024) is pretrained on extensive datasets, and it is transferred to execute specific tasks through appropriate fine-tuning (Wang L et al., 2025).

In the following section, we propose to transplant this "pretraining+fine-tuning" paradigm to quantitative investment. Specifically, in the pretraining stage, we build a universal foundational model using financial data from different exchanges, different instruments, and different frequencies to discover transferable trading patterns and investment logics. Then the following quantitative strategy research might be reconceived as a fine-tuning task tailored to specific strategy requirements and investment scenarios. Such a paradigm shift could dramatically increase research efficiency in the field.

3.1 Universal modeling for LIM

Universality of LIM should encompass at least the following three aspects:

1. Cross-instrument universality. Given quote data, can a machine-mined pattern for predicting stock trends be applied to predict trends in futures or bonds? Logically, many trading patterns reflect traders' intentions and behaviors, and it is natural for some common patterns to be shared across various instruments. Empirically, many technical indicators or price/volume factors are useful not only

for stock prediction but also for bonds, futures, and even cryptocurrencies. This suggests the feasibility of training a general-purpose upstream model with data from various instruments and fine-tuning this model with specific data for each instrument.

2. Cross-exchange universality. Stock trading across different exchanges may exhibit common patterns or signals, especially for technical indicators or strategies based on quote data (and sometimes news data). This observation motivates the development of a "universal" model using data from multiple exchanges, which can then be applied to trade equities in specific markets. Similarly, many other instruments (e.g., futures and bonds) share cross-exchange patterns, making them suitable for pretraining the foundational quant model.

3. Cross-frequency universality. Patterns often persist across different data frequencies, such as 1-s, 15-s, 1-min, 20-min, 1-h, and 1-d candlesticks. Training on data from the same instrument across various frequencies can significantly enhance the sample size, which is crucial for improving the performance of deep learning models.

As shown in Fig. 6, the architecture follows a typical "pretraining+fine-tuning" structure. First, the upstream model acts as a generative foundation for quantitative finance, simplifying strategy formulation into a unified framework akin to a "next-token prediction" problem (Li et al., 2024). Utilizing self-supervised learning (Bengio et al., 2013), the model efficiently learns representations from various financial data across different instruments, exchanges, and frequencies, capturing nuanced market patterns and relationships. This predictive approach streamlines strategy development by transforming traditional task-specific modeling into a more generalized sequence prediction problem. Consequently, the model can infer future market conditions based on historical data, similar to how language models predict the next word in a sentence, thereby unlocking new potential for quant strategy research, algorithmic trading, risk assessment, and portfolio management in a more automated and scalable manner. Second, the downstream model fine-tunes the upstream model according to specific task requirements to develop quantitative trading

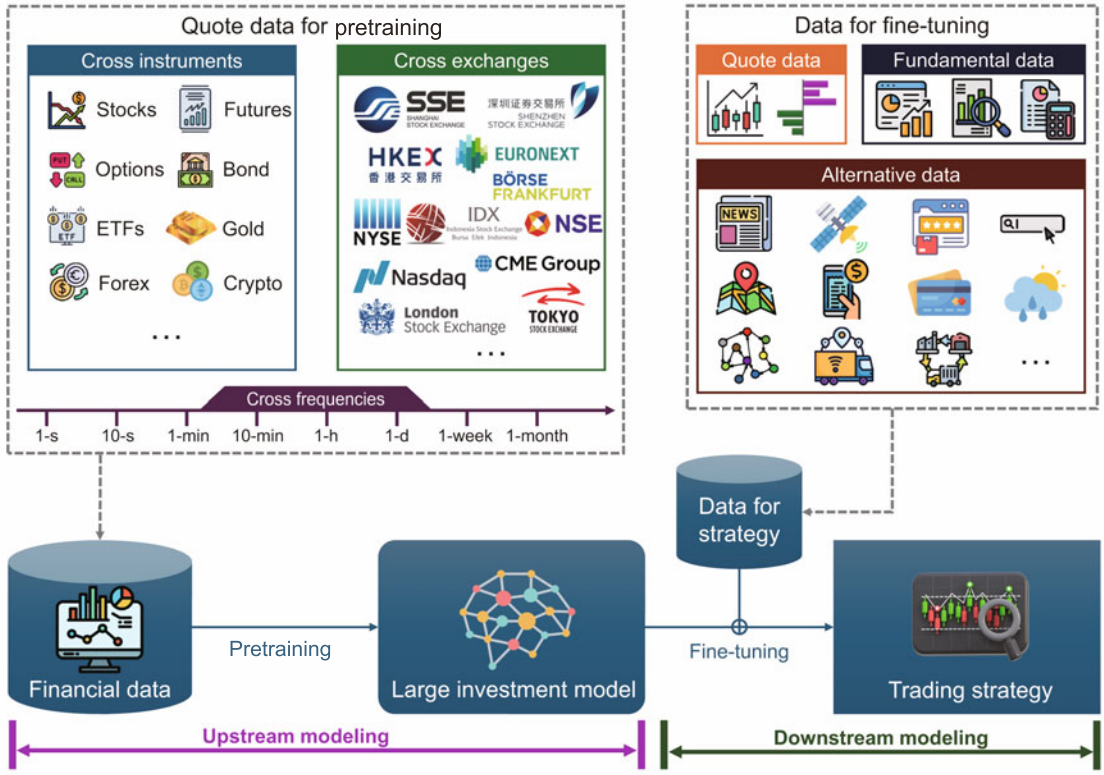

Fig. 6 Workflow of the large investment model (LIM)

strategies, including momentum strategies, mean-reversion strategies, pairs trading strategies, triangular arbitrage strategies, calendar spread arbitrage, and cross-sectional hedging strategies. Given the diverse specifications and configurations of different tasks, downstream modeling employs a range of approaches.

3.2 Practical issues for building LIM

Given the LIM framework, we have to consider several practical problems in reality:

1. Data quality issue. The low quality of financial data caused by missing values, noise, typos, etc. may negatively affect the performance of quantitative strategies. This issue is more serious in fundamental data and alternative data; for example, a fake news story may mislead the strategy decision and lose money. On the contrary, quote data, including prices, volumes, and LOBs, usually have much higher quality, and they are universal data types among different exchanges and different instruments.

2. Backbone of the foundation model. Compared with fundamental data and alternative data, quote data have better completeness, consistency, and universality across various exchanges and instruments. Therefore, we choose to use quote data to build the backbone of the foundation model and use its embedded output to enhance downstream prediction tasks.

3. Universality in building the foundation model. To maximize the universality of the upstream model, we define it as a single time-series processor which is trained using only the time-series information of each single instrument and no cross-instrument information. Therefore, strategies that depend on multiple instruments, such as pairwise arbitrage trading (Krauss, 2017) and cross-sectional stock long–short trading (Fama and French, 1992), will be modeled during the downstream fine-tuning phase.

4. Model transferability issue. The transferability of the models between markets is quite different. Empirically, for example, a model trained with Japan stock data is easier to transfer to predict Korea stock markets rather than US stock markets. Therefore, an important preprocessing task is to determine the markets that are more transferable and whose data are appropriate for training the foundation model.

5. Model update issue. Because the foundation model is trained over time, the model needs to be updated on time. The update frequency depends on a number of considerations, for example, the sufficiency of computing power, the period of prediction horizon, and the movement of market structure. Empirically, a weekly update of the foundation model is sufficient for most downstream tasks.

6. Risk management issue. Different tasks require different risk management principles and methods, and therefore we leave the risk management issues in downstream tasks. Take the stock cross-sectional selection strategy as an example. The Barra risk factor model (Sheikh, 1996) by Morgan Stanley Capital International (MSCI) is regularly used to control risk exposures from various market styles and industries. On the other hand, for commodity trading advisor (CTA) strategies, an appropriate algorithmic drawdown controller is needed to control the risk of loss.

3.3 An illustrative example for LIM model design

To demonstrate the LIM framework, we present a cross-market modeling case study that illustrates its architecture design in practice. In the upstream phase, we use Transformer (Vaswani et al., 2017) to develop a foundation model trained on stock data from both Chinese (Shanghai and Shenzhen exchanges) and US markets (NYSE and Nasdaq). This model is subsequently adapted for predicting price movements in China's A-share market through downstream transfer learning.

Fig. 7 shows the LIM architecture. For both markets, we process stock time-series data using 5-min candlestick charts within 150-min rolling windows. From these, we extract $M = m \times 30$ features, where m represents the number of basic features obtained from each bar (including information about log-returns, trading volumes, and turnovers etc.). The model predicts returns across multiple horizons (e.g., 5-min, 15-min, and 30-min), resulting in multi-dimensional prediction targets. We employ a composite loss function combining IC optimization (Xia and Simonian, 2021) with mean-squared error (MSE) minimization (Hastie et al., 2009). The dataset spans from 2015 to 2024, partitioned into training (2015–2022), validation (2023), and test

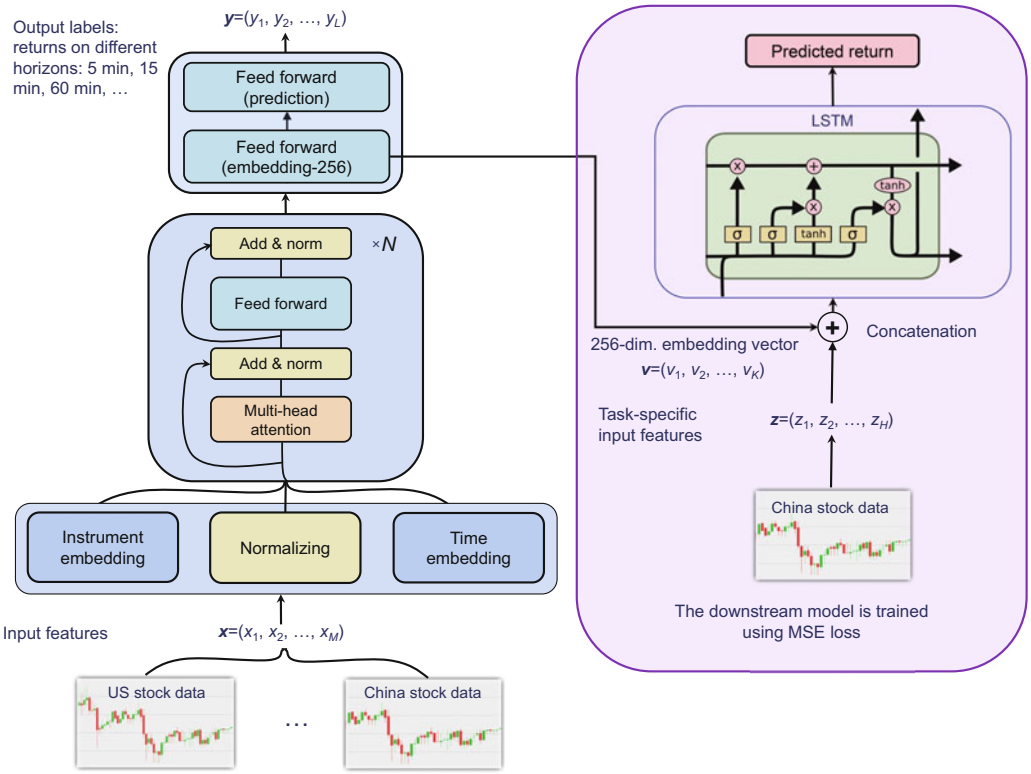

Fig. 7 A model design example for the large investment model (LIM). The diagram on the left is the upstream model (pretrained using stock data from China, the United States, and other markets), whose main body is a Transformer network. The diagram on the right is the downstream model, which accepts the embedding vector from the upstream model and concatenates it with the feature vector extracted from the downstream task (predicting the returns of the stocks in the China A-share market) to train a downstream LSTM model for A-share stock return prediction

(2024) periods.

The upstream Transformer architecture comprises N attention layers followed by two sequential feedforward modules. The first module generates a 256-dimensional embedding vector for downstream transfer, while the second produces the final predictions $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L)$. During downstream fine-tuning, this pretrained embedding is concatenated with China A-share market features to train an LSTM network (Hochreiter and Schmidhuber, 1997) specifically for 30-min return prediction.

We evaluate model performance using the time-series information coefficient (TSIC), defined as $\text{TSIC} = \text{Corr}(\hat{r}, r)$, where $\hat{r}$ denotes predicted future 30-min returns and r represents actual returns. Higher TSIC value indicates superior prediction accuracy. As shown in Table 1, the LIM-style pretrained model achieves a TSIC of 0.088, representing an 18.9% improvement over the baseline model's 0.074 (trained solely on China A-share data). This performance enhancement suggests that the cross-market patterns learned from US market data contain transferable predictive signals for China A-share movements. The results demonstrate the value of multi-market pretraining in improving financial time-series forecasting accuracy.

Table 1 Time-series information coefficient (TSIC) between two modeling approaches

Model	TSIC
Baseline	0.074
Pretrain+Fine-tune	0.088

The baseline model was trained exclusively on China A-share market data, while the pretrain+fine-tune model was first pretrained on combined China and US market data before being fine-tuned on China A-share. These two models were evaluated on the same China A-share test set

4 Upstream foundation model

As illustrated in Fig. 8, the upstream modeling focuses on developing a universal foundation model for quantitative investment. The goal of the upstream pretraining foundation model is to be as general as possible, to address a broad spectrum of financial time-series prediction problems.

4.1 Problem formulation

We formulate the foundation model in the upstream pretraining stage as follows. Suppose that we have a dataset of M multivariate time series $\mathcal{D}=\{\boldsymbol{X}^{(m)}\}_{m=1}^{M}$, where each multivariate time series $\boldsymbol{X}^{(m)}\in\mathbb{R}^{T^{(m)}\times p}$ has a length $T^{(m)}$ and dimension p. For each time point t $(1\le t\le T)$, we define two sliding windows: a look-back context window of length L_X that defines the input $\boldsymbol{X}_{t-L_X+1:t}$, and a look-forward horizon window of length L_Y that defines the output $\boldsymbol{X}_{t+1:t+L_Y}$. The foundation model is a function $f:\mathbb{R}^{L_X\times p}\to\mathbb{R}^{L_Y\times p}$. Given parameters $\boldsymbol{\Theta}$, we aim to satisfy the relationship $\boldsymbol{X}_{t+1:t+L_Y}\approx f(\boldsymbol{X}_{t-L_X+1:t}\mid\boldsymbol{\Theta})$.

To estimate the mapping function f, various deep learning models can be used by minimizing the loss function:

$$\sum_m\sum_t L(\boldsymbol{X}_{t+1:t+L_Y},f(\boldsymbol{X}_{t-L_X+1:t}\mid\boldsymbol{\Theta})),\qquad(1)$$

where $L(\cdot,\cdot)$ is a predefined distance metric, such as MSE or cross-entropy loss (Mao et al., 2023). Under

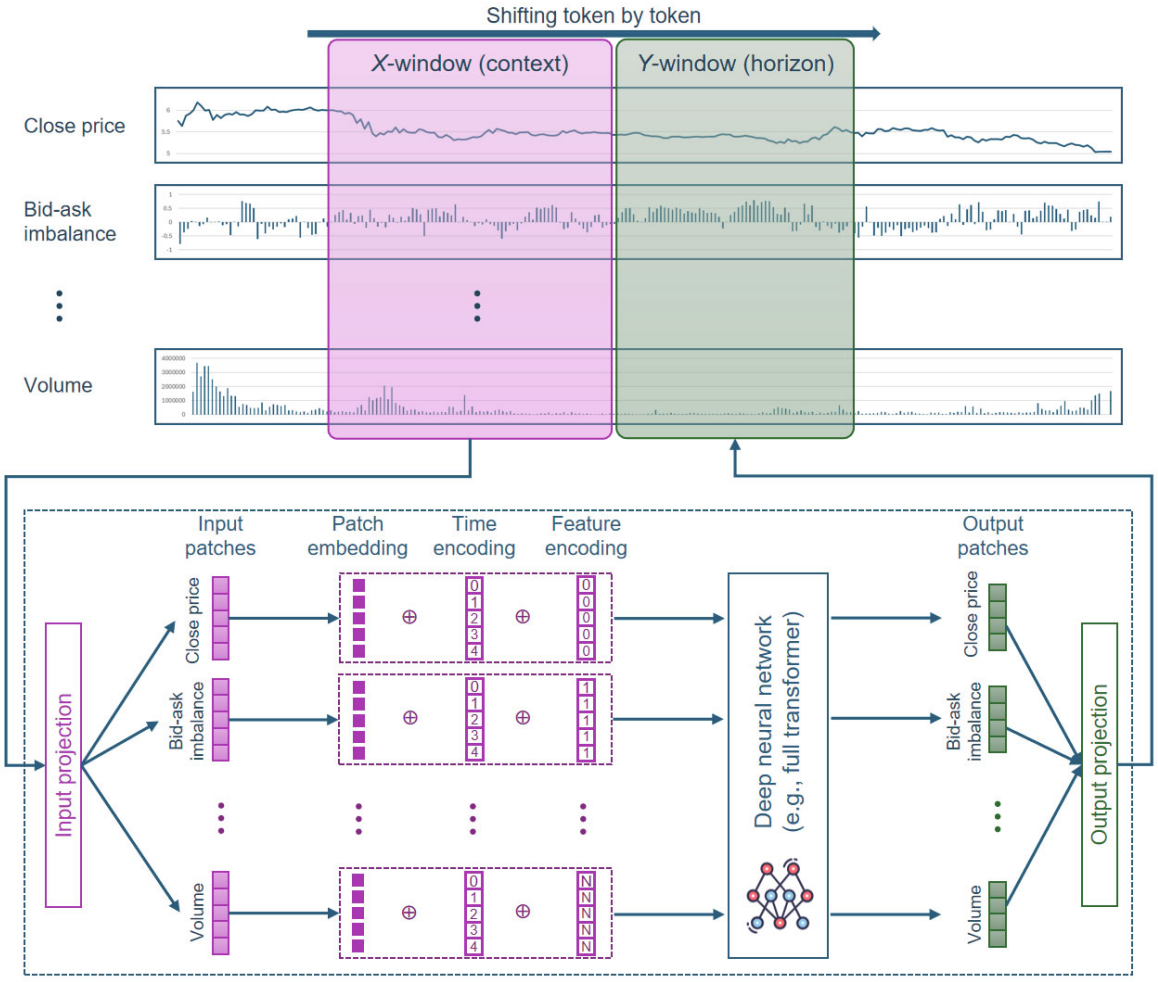

Fig. 8 Illustration of the workflow for the upstream foundation model. Financial quote data (price/volume and LOB) are used to construct the backbone of the foundation model, while other data types, including fundamental and alternative data, are integrated into the backbone aligned with the time axis and investment instruments

the assumptions of continuity and differentiability, this loss function can be differentiated to derive the gradient, which is used to optimize the parameters during the training of deep neural networks.

4.2 Design of the foundation model

Fig. 8 illustrates the construction of an upstream foundation model for LIM. This model is designed to predict future tokens within a Y-window (horizon) based on data from an X-window (context) that covers a fixed historical period. The X-window data are fed into the modeling module and serve as the input for the backbone deep learning model. This model is trained on financial quote time-series data, incorporating various variables (meta-features) such as closing price, bid-ask imbalance (Brown P et al., 1997), returns, and trading volume, to predict the same set of variables within the Y-window. To enhance computational efficiency, time-series segmentation techniques such as patching (Nie et al., 2023) are applied within the windows. Additionally, patch masking strategies (Das A et al., 2024) are employed to improve the quality of self-supervised learning and increase the flexibility of window length during model training.

For each variable in the time series, the input data from the X-window are first transformed into input patch vectors via an input projection, which then generates patch-embedding vectors. These embeddings are concatenated with a time-encoding vector to capture temporal information and a feature-encoding vector to identify the specific variable being used. After processing through a deep neural network (e.g., a Transformer), the model outputs patches corresponding to each variable. These outputs are then merged and converted back to the original time-series granularity through output projection, ultimately generating predictions for the variables in the Y-window.

In addition, significant differences in trading rules and transaction costs across exchanges exist. For example, some stock exchanges operate on a $T+1$ trading basis (stocks bought today cannot be sold before the next trading day), whereas others use a $T+0$ system, allowing for same-day trading. These differences affect the timing of cash flows, market patterns, and overall strategy design and execution. Because the LIM foundation model is expected to learn common market patterns, these differences across exchanges are ignored, treating the problem as a pure time-series prediction task.

5 Downstream task model

The downstream workflow bridges the foundation model from the upstream process to the final strategy development task. Unlike the foundation model, which primarily relies on quote data, downstream modeling can incorporate a wide variety of task-specific data sources, including news, supply chain information, satellite imagery, and earnings call transcripts. These diverse data types can be categorized into graph data, textual data, image data, numerical data, audio data, and video data. To effectively utilize these additional inputs, we employ specialized embedding techniques tailored to each data type, enabling the model to integrate and leverage the unique information contained within these varied structures (Fig. 9).

5.1 Data preprocessing

Handling diverse data types is crucial for developing robust quant models. Fundamental data, such as financial statements, and alternative data, such as social media sentiment and satellite imagery, often exhibit irregularities in their time series. This irregularity poses significant challenges for data preprocessing, necessitating a methodical approach for aligning data accurately with corresponding time points and financial instruments.

Proper alignment (Fig. 10) is essential to ensure that the input data used for model training and evaluation are coherent, consistent, and reflective of true market conditions. First, a fundamental step in data preprocessing is temporal alignment. Since fundamental and alternative data points do not always coincide with regular intervals, it is necessary to synchronize these data points to a common timeline that matches the trading dates or specific events relevant to the investment strategy. Techniques such as interpolation and nearest-neighbor methods (Ahlberg et al., 1967) can be employed to estimate missing values and align the data with the appropriate time stamps. This alignment ensures that the models receive continuous and accurate inputs, thereby enhancing their predictive accuracy and reliability.

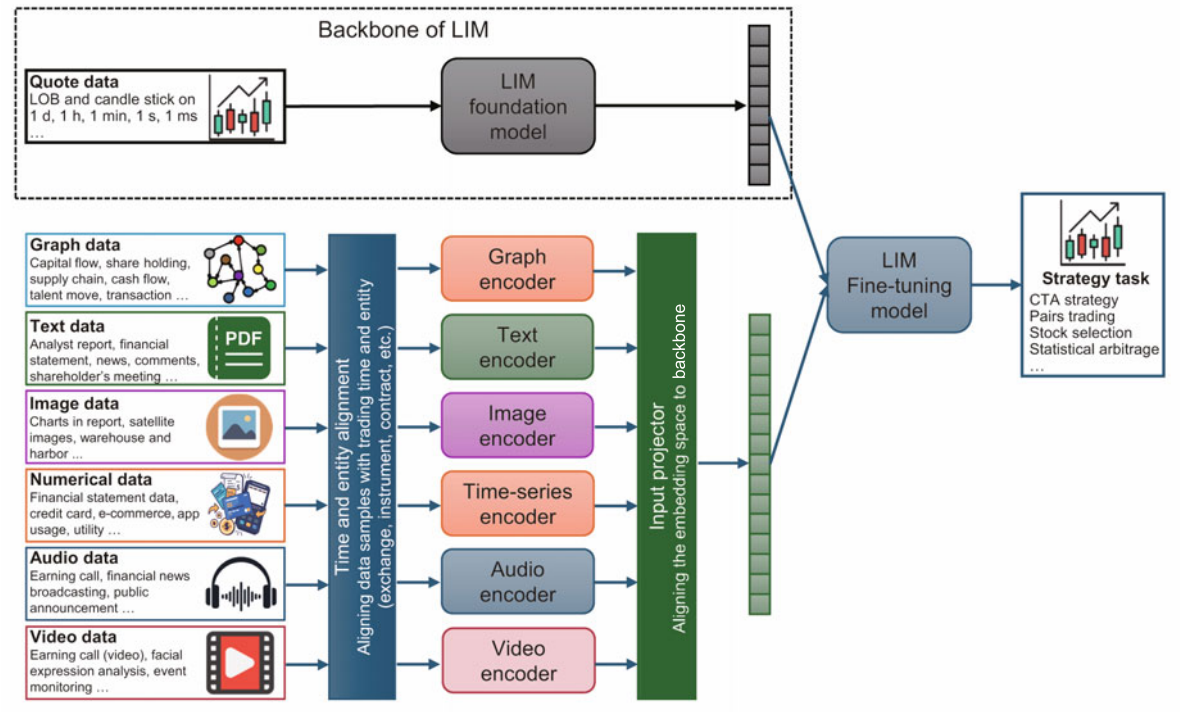

Fig. 9 Downstream workflow for the large investment model (LIM)

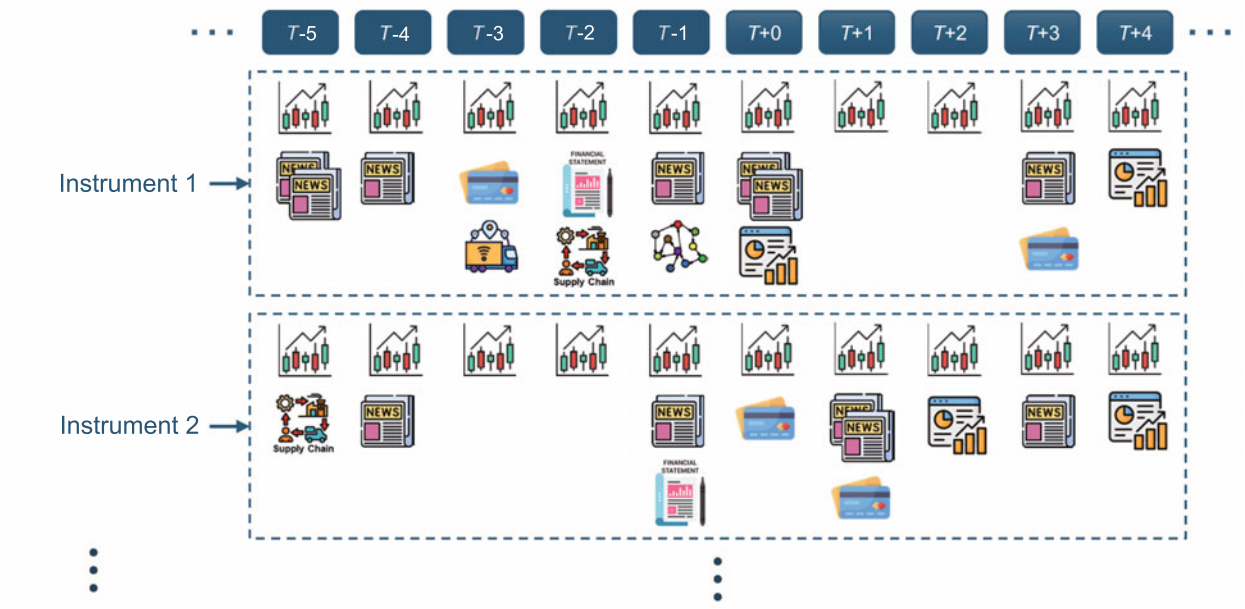

Fig. 10 Aligning diverse data with time patch and financial entities

Second, instrument alignment is equally critical in the preprocessing phase. Given that different financial instruments (e.g., stocks, bonds, and derivatives) may have unique characteristics and response patterns to various data inputs, aligning data to the correct instrument is imperative. This involves mapping the fundamental and alternative data to the specific instruments to which they relate, ensuring that each data point is correctly attributed. For instance, corporate earnings reports must be matched to the

corresponding company's stock, while industry-wide metrics should be appropriately linked to all relevant securities within that industry.

Furthermore, aligning data across time points and instruments involves winsorization (Dixon W and Yuen, 1974) and standardization processes. Winsorization mitigates the impact of outliers, whereas standardization ensures that all data inputs belong to a consistent scale, facilitating more efficient training of machine learning models. By incorporating these preprocessing steps, the data fed into deep neural network models become more robust, and this helps reduce noise and improve the overall quality of the predictions.

5.2 Model fine-tuning

Adapting the foundation model to meet current strategy development demands involves employing several common fine-tuning methods appropriate for quantitative investment. Transfer learning (Pan and Yang, 2010) is a pivotal technique where the pretrained model is fine-tuned on a new, specific dataset, allowing it to retain its learned knowledge while becoming more specialized. Feature-based transfer learning (Daumé, 2007) adjusts only the last few layers, leveraging the previously learned features, while fine-tuning entire models (Yosinski et al., 2014) retrain the entire network to adapt to the new domain. Layer-wise fine-tuning (Howard and Ruder, 2018) is a technique where different layers of the model are fine-tuned at different rates, typically starting from the last layers and progressively fine-tuning earlier layers. Parameter-efficient tuning methods, such as low-rank adaptation (LoRA) (Hu EJ et al., 2021) and adapter modules (Pfeiffer et al., 2021), introduce a small number of trainable parameters to existing pretrained models, allowing for efficient fine-tuning with fewer resources. These algorithms are chosen based on the nature of the task, available data, and computational resources, balancing the need for adaptation with the risk of overfitting. For instance, when working with smaller pretrained models (under 1 billion parameters), we recommend full fine-tuning (retraining the entire model on downstream task data) to achieve optimal performance. Conversely, for large-scale pretrained models (with hundreds of billions of parameters), we suggest either feature-based transfer learning (adjusting only the final layers) or parameter-efficient methods such as LoRA. These approaches significantly improve training efficiency while only marginally compromising prediction accuracy.

5.3 Various types of downstream tasks

Quantitative investment encompasses a broad range of strategy types. Given that the upstream foundation model primarily serves as a predictor for single time-series data, the downstream process becomes crucial for strategies that involve multiple instrument time-series. These include strategies such as cross-sectional stock selection, pairs trading, and various complex arbitrage approaches. Below, we outline several typical strategy scenarios and describe the corresponding downstream processing tasks:

1. Fundamental investing. Fundamental investing, which often involves low-frequency trading, relies heavily on fundamental data sources such as analysts' reports, financial statements, news articles, and other alternative data related to company performance and operations. In the downstream process, these data types are combined with the embeddings generated by the pretrained model to fine-tune a new prediction model focused on fundamental analysis. Due to the typically small sample size in low-frequency trading, the downstream model is often limited to predicting the next alpha signal.

2. Statistical arbitrage. Statistical arbitrage (Krauss, 2017) strategies involve trading two or more historically correlated assets based on deviations from their mean or expected relationship. For instance, in pairs trading, when one asset outperforms its counterpart, the strategy may involve selling or shorting the overvalued asset while buying or going long on the undervalued asset, with the expectation that the spread will revert to its historical average. In this context, the upstream foundation model embeds the assets used in pairs trading into latent vectors. These vectors are then processed by the downstream model to predict optimal trading times.

3. Lead–lag strategy. The lead–lag strategy (Li et al., 2022) involves trading two assets where one asset (the "leading" asset) is anticipated to influence the performance of the other (the "lagging" asset). Unlike pairs trading, where assets are traded

in opposite directions, the lead–lag strategy involves trading only the lagging asset based on the trend of the leading asset. In this scenario, the downstream model is fine-tuned to take embeddings of both the lead and lag time series as input, and it outputs predictions for the lagging series.

4. Cross-sectional strategy. Cross-sectional strategies (Engelberg et al., 2023) differ from time-series approaches in that they involve trading a broad universe of assets simultaneously based on predicted alphas for the same time horizon. Common cross-sectional strategies include stock selection and long–short hedging. During the downstream process, the entire cross-section of assets is inputted into the fine-tuning model as a single sample. Multiple cross-sectional samples from different time points are used to train the downstream model, ultimately guiding the selection of stocks for buy/sell or long/short trades based on the predicted horizons.

6 System architecture for LIM

This section outlines the construction of a real-world system founded on the LIM methodological framework. This comprehensive system supports the entire modeling pipeline, including the computing infrastructure, data computation and storage, foundation modeling and management, automated strategy modeling, human–AI interaction agents, and a low-latency trading system (Fig. 11).

6.1 Computing and data infrastructure

Building large-scale investment models requires the integration of high-performance computing (HPC) platforms to manage the complexity and volume of data. These platforms facilitate efficient training and execution of models, ensuring scalability to accommodate growing data volumes and computational demands (Dempster et al., 2018). The architecture of HPC systems can be specially optimized for financial time-series modeling, enhancing the performance of deep learning models applied to these data types.

An effective and reliable data system is crucial for deploying LIM. This system must support a variety of database types to meet different data storage and retrieval needs:

1. SQL databases manage relational data such as candlestick data, financial statement records, and transaction data to ensure robust data integrity and complex query capabilities (Stonebraker, 2010).

2. Graph databases are excellent for handling data with complex relationships and interconnections, which are useful for analyzing supply chain networks and stock relationships (Angles and Gutierrez, 2008).

3. NoSQL databases are suited for unstructured and semi-structured data, like social media feeds and news articles, offering flexibility and scalability (Han J et al., 2011).

4. Time-series databases specialize in managing temporal data, which are crucial for tracking financial market data and economic indicators (Dunning et al., 2014).

5. Vector databases store high-dimensional embedding vectors that characterize various data representations and manage metadata from diverse alternative data sources (Xie et al., 2023).

To complement this diverse data system, constructing a high-performance data computation system is essential for accelerating data preprocessing tasks. Utilizing distributed computing frameworks such as Apache Spark (Salloum et al., 2016) allows for parallel processing of large datasets, significantly reducing the time required for data cleaning, transformation, and integration. In-memory computing technologies such as Apache Ignite (Stan et al., 2019) and Redis (Das V, 2015) enhance processing speeds by storing data in RAM for quick access and manipulation. For real-time trading, stream processing frameworks like Apache Flink (Carbone et al., 2015) provide continuous ingestion and processing of real-time data streams, ensuring timely decision-making for investment models.

Building a highly reliable data system supports large-scale investment models by ensuring data integrity, availability, and security. Distributed storage solutions such as the Hadoop distributed file system (HDFS) (Shvachko et al., 2010) offer scalable and fault-tolerant storage that distributes data across multiple nodes to ensure redundancy and high availability. Efficient computing is further achieved through algorithm optimization and advanced hardware, enhancing computational efficiency and enabling faster, more accurate predictions. Data robustness is maintained through rigorous validation

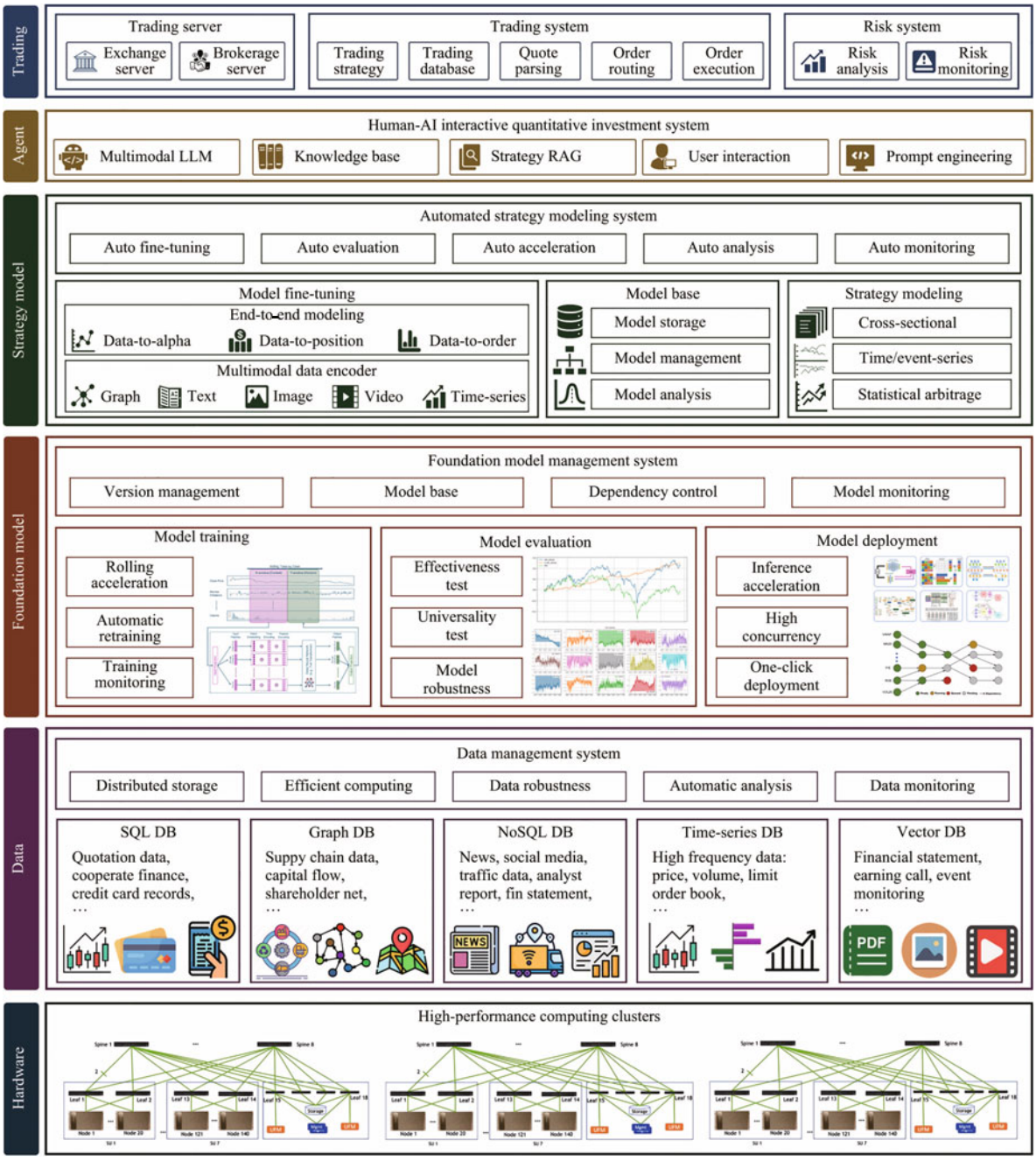

Fig. 11 Architecture design for an LIM system platform, which is composed of a hardware layer, data layer, foundation model layer, strategy model layer, agent layer, and trading layer, with each layer consisting of a number of system modules

and cleansing procedures, alongside redundant storage systems to mitigate the risks of data corruption and loss. Automated data analysis and comprehensive computation monitoring with tools like Prometheus and Grafana provide real-time insights into system performance, facilitating proactive management and optimization of computing resources (Saputra et al., 2024).

6.2 Systems for the foundation model

The core of an LIM system begins with the construction of a comprehensive module for foundation modeling training. This module integrates several

advanced technical features to ensure efficiency and accuracy. A pivotal feature is rolling training acceleration, achieved through incremental learning or continuous learning techniques (Wang L et al., 2024). This allows the model to continually update itself with new data without the need for retraining from scratch, thus significantly reducing both training time and computational resource demands. Automatic retraining is implemented to keep the model current by periodically integrating the latest data, thus maintaining the model's performance and relevance over time. Moreover, training monitoring is crucial for managing the training process, involving real-time tracking of metrics such as training loss and validation accuracy, to facilitate the early detection of issues and enable timely adjustments (Goodfellow et al., 2016).

Once the foundation model is trained, a rigorous evaluation process is necessary to validate its effectiveness and reliability. This involves the construction of a dedicated system module for comprehensive model evaluation, encompassing several key tests:

1. Effectiveness test. This employs the backtest to assess the model's performance using historical data to simulate real-world scenarios, helping identify potential weaknesses and areas for improvement.

2. Universality test. This examines the model's applicability across different financial instruments and exchanges, ensuring that it can generalize well across various assets and market conditions.

3. Robustness test. This test evaluates the stability of the model's performance under different market conditions, including stress testing (Sorge, 2004) scenarios, to ensure effectiveness even in volatile or adverse markets.

6.3 Automated strategy modeling

Addressing the diverse demands of various strategy tasks is crucial in AI quantitative investment. Building an automated strategy generator that can fine-tune, evaluate, analyze, and monitor new strategy models significantly enhances the efficiency of strategy research. This generator automates these processes, accelerating the development of new strategies and ensuring consistency and precision in model adjustments and evaluations, thus swiftly adapting to new market conditions and uncovering innovative trading strategies.

Unlike foundation modeling, which may rely on more homogeneous data sets, the fine-tuning module implements modeling using a broader variety of data types. By integrating diversified data sources, the module can enhance the model performance, leading to more accurate and robust trading strategies (Bengio et al., 2013).

Models, once fine-tuned, are transformed by the strategy modeling module into actionable trading strategies for quantitative investment, including various types such as cross-sectional trading, time-series trading, event-driven trading, and statistical arbitrage trading. This transformation ensures that theoretical model improvements are translated into practical trading gains.

6.4 Agent system

Enhancing human–AI interaction and model explainability is important for improving research efficiency. The LIM agent system uses natural language processing to ensure that the models are both high-performing and explainable, which is essential for building user trust and improving decision-making. Techniques such as multimodal LLMs (Yin et al., 2024), knowledge graph (Zhong et al., 2023), and retrieval-augmented generation (RAG) (Lewis et al., 2020) enhance the system's functionality and user experience by processing complex queries and providing contextually relevant interactions.

6.5 Trading system

Effective deployment of AI-driven quantitative investment strategies necessitates integration with brokers or exchanges for real trading. This integration involves establishing secure and reliable communication channels with trading platforms, enabling real-time order-sending, market data receipt, and trade execution monitoring. A comprehensive trading module within the system is essential for efficiently executing these strategies, including components like trading strategy formulation, trading database management, quote parsing, order routing, and order execution, optimized to minimize trading latency and maximize execution quality (Aldridge, 2013).

7 Future research directions

Although we have introduced the basic framework for the LIM in this article, there are still a

series of advanced algorithms and technical issues that remain unresolved and require further in-depth research. Below, we highlight several potential research directions for further exploration:

1. End-to-end risk modeling. In quantitative investment, risk management is at least as crucial as sound strategies and substantial returns. Traditional multifactor risk models like the Barra model attempt to decompose portfolio returns and volatility into linear combinations of various risk factors such as size, beta, leverage, and liquidity. These models' performance depends on the effectiveness of the risk factors used. To address the limitations of multifactor risk models, it is worthwhile to study end-to-end deep learning risk models that can be used to neutralize alpha models and reduce portfolio volatility.

2. Market simulation. Backtesting simulations are widely used in academia and industry to evaluate new strategies before real trading in the market. However, backtests ignore market impact on actual trading cost, resulting in discrepancies between backtests and real trading. Therefore, it is worthwhile to develop a market simulator for testing new strategies and for estimating potential market impact at varying asset volumes. Existing research on market simulation focuses on micro-markets, like LOB generators (Coletta et al., 2023; Nagy et al., 2023). A more general research problem for LIM is how to build a financial world model that simulates both micro- and macro-markets, with a full spectrum of data resolutions and types.

3. Model on multiple granularities. Trading can occur at various granularities, from millisecond-level high-frequency trading to year-long low-frequency investment. A successful quantitative trading strategy may involve multi-horizon predictions that are integrated to enhance final trade decisions. Researching new deep neural network frameworks for multi-scale prediction using multi-granularity data is valuable for quantitative investment.

4. LIM with multiple backbones. The LIM framework uses financial quote data to build the backbone model. However, as the horizon extends to low-frequency strategies, textual data like news and financial reports become increasingly important and can be used to build LLMs that serve as a secondary backbone model. It is interesting to study how to integrate these two backbones within the same LIM framework.

5. More comprehensive multimodal LIM. With more alternative data being used by financial institutions, the LIM architecture must be flexible enough to accommodate new data types, like audio and video data for important conference calls. Additionally, LIM uses a plug-and-play mechanism to support automatically aligning new types of time-series data through appropriate embedding combinations.

6. Extending LIM with agents. The current LIM shares the same limitation as LLMs in that data and knowledge cannot be updated in real time because pretraining can take several weeks. The agent framework can address LIM deficiencies by incorporating real-time updated knowledge bases, search engine information, and other data sources. It can also enhance LIM's reasoning power through multi-agent debating, reflection, and other agent techniques.

7. Improving the explainability of LIM. Understanding investment strategies and decisions is essential for any investor. Unlike traditional explainable machine learning techniques (Yang et al., 2023) focused on feature or attention importance, LLM techniques offer a new approach for modeling explainability through logical reasoning and natural language interaction. It is particularly valuable to research how to improve LIM techniques to project latent investment logic embedded in the "black-box" deep learning model into human-understandable natural language and illustrative charts/images.

8. Inference acceleration for LIM. The computational demands of deep learning models, especially during inference, can be substantial for complex financial models that require real-time decision-making based on large data volumes. To reduce computation and execution latency and increase real-time data throughput, standard inference acceleration techniques like model quantization (Jacob et al., 2018), pruning (Han S et al., 2015), and efficient deployment strategies can be evaluated and selected. Additionally, researchers are encouraged to develop new acceleration algorithms tailored to financial scenarios. For example, the temporal nature of financial data may offer opportunities to accelerate attention computation in Transformer-like neural networks.

9. New architecture for LIM. Researching new neural network architectures for financial time series

is necessary to enhance effectiveness, efficiency, and robustness. The new end-to-end, unified architecture should handle irregularly sampled time-series data, various data granularities, diversified data structures and knowledge attributes, and extended in-context learning abilities. Efforts should focus on further improving LIM's forecasting performance for different investment strategies.

8 Discussion and conclusions

LIM offers a transformative approach to quantitative investment, fundamentally reshaping how financial models are developed, trained, and applied across diverse market environments. First, LIM serves as a knowledge transfer process, enabling the model to learn from global market data and apply this knowledge to local markets. This transfer involves intelligent adaptation that leverages broad patterns observed across various financial environments to improve predictive accuracy and robustness in specific contexts. Second, LIM functions as an advanced data augmentation process, integrating diverse financial data sources, including equities, futures, and commodities, to create a richly diversified training environment. This allows LIM to capture complex interrelationships and patterns that might be missed in more narrowly focused models, making it particularly valuable in today's complex financial markets. Third, LIM significantly reduces the cost and improves the efficiency of modeling. Once the foundation model is trained on global data, it can be fine-tuned for specific tasks with minimal additional effort, eliminating the need for expensive and time-consuming training from scratch. Finally, LIM provides a unique opportunity to uncover potential correlations or causal relationships between instruments across different markets and asset types (e.g., lead–lag effects).

Given its advantages, LIM also presents significant technical challenges that must be addressed to fully realize its potential in quantitative investment. First, a major challenge is the high maintenance cost, as the foundation model requires frequent retraining to stay effective and up-to-date. This retraining, which might be needed daily, weekly, or monthly depending on market dynamics and strategy frequency, is computationally intensive and demands substantial infrastructure, increasing operational costs. Managing these costs while maintaining model accuracy and reliability requires sophisticated resource management. Second, selecting appropriate exchanges, instruments, and data frequencies for training is critical. The diversity of financial markets means that not all instruments contribute equally to the model's performance, and the heterogeneity of data can introduce noise or even harmful patterns if not carefully curated. Significant effort is needed to identify the most beneficial data combinations. Third, integrating alternative data sources adds another layer of complexity. While alternative data can enhance the predictive power of downstream models, determining which types (such as social media sentiment and satellite imagery) are useful requires extensive research and validation. Finally, the current LIM implementation is most effective for high- or medium-frequency trading strategies, as it primarily relies on quote data. This limits its applicability to low-frequency strategies like value investing and global macro investing, which rely more on fundamental data such as financial statements and economic indicators. Extending LIM's applicability to these areas will require significant research and development to integrate fundamental and alternative data into the foundation model.

In conclusion, this article introduces LIM, a universal and scalable framework designed to advance quantitative investment research. By integrating diverse data sources and applying them across various market contexts, LIM can accelerate the development of more sophisticated and effective investment strategies, potentially leading to improved investment outcomes.

Acknowledgments

The authors would like to thank Saizhuo WANG, Sida LIN, Haohan ZHANG, Chengjin XU, Yiyan QI, Zhouchi LIN, Hang YUAN, Bokai CAO, and Xinyi LIN for their comments, suggestions, and support.

Contributors

Jian GUO and Heung-Yeung SHUM jointly designed the LIM framework. Jian GUO conducted detailed design about modeling and drafted the paper. Jian GUO and Heung-Yeung SHUM revised and finalized the paper.

Conflict of interest

Jian GUO is a guest editor of the Special Feature on Theories and Applications of Financial Large Models of *Frontiers of Information Technology & Electronic Engineering* (*FITEE*). Heung-Yeung SHUM is an editorial board member of *FITEE*. They were not involved with the peer-review process of this paper. Both authors declare that they have no conflict of interest.

Open access

References

Ahlberg JH, Nilson EN, Walsh JL, 1967. The Theory of Splines and Their Applications. Academic Press, New York, NY, USA.

Ait-Sahalia Y, Saglam M, 2024. High frequency market making: the role of speed. *J Econometr*, 239(2):105421. https://doi.org/10.1016/j.jeconom.2022.12.015

Aldridge I, 2013. High-Frequency Trading: a Practical Guide to Algorithmic Strategies and Trading Systems (2nd Ed.). Wiley, Hoboken, USA.

Angles R, Gutierrez C, 2008. Survey of graph database models. *ACM Comput Surv*, 40(1):1-39. https://doi.org/10.1145/1322432.1322433

Bengio Y, Courville A, Vincent P, 2013. Representation learning: a review and new perspectives. *IEEE Trans Patt Anal Mach Intell*, 35(8):1798-1828. https://doi.org/10.1109/TPAMI.2013.50

Black F, Scholes M, 1973. The pricing of options and corporate liabilities. *J Pol Econ*, 81(3):637-654. https://doi.org/10.1086/260062

Bodie Z, Kane A, Marcus AJ, 2014. Investments (10th Ed.). McGraw-Hill Education, New York, USA.

Brown P, Walsh D, Yuen A, 1997. The interaction between order imbalance and stock price. *Pacif-Bas Fin J*, 5(5):539-557. https://doi.org/10.1016/S0927-538X(97)00019-X

Brown T, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33:1877-1901.

Cao SS, Jiang W, Lei LG, et al., 2024. Applied AI for finance and accounting: alternative data and opportunities. *Pacif-Bas Fin J*, 84:102307. https://doi.org/10.1016/j.pacfin.2024.102307

Carbone P, Katsifodimos A, Ewen S, et al., 2015. Apache Flink: stream and batch processing in a single engine. *IEEE Datab Eng Bull*, 36:28-33.

Chan KC, 1988. On the contrarian investment strategy. *J Bus*, 61(2):147-163.

Chen Q, Francis J, Jiang W, 2005. Investor learning about analyst predictive ability. *J Account Econ*, 39(1):3-24.

Chen T, Chen W, Du L, 2021. An empirical study of financial factor mining based on gene expression programming. 4th Int Conf on Advanced Electronic Materials, Computers and Software Engineering, p.1113-1117. https://doi.org/10.1109/AEMCSE51986.2021.00228

Chincarini L, Kim D, 2022. Quantitative Equity Portfolio Management: an Active Approach to Portfolio Construction and Management (2nd Ed.). McGraw-Hill Education, New York, USA.

Cho P, Park JH, Song JW, 2021. Equity research report-driven investment strategy in Korea using binary classification on stock price direction. *IEEE Access*, 9:46364-46373.

Coletta A, Jerome J, Savani R, et al., 2023. Conditional generators for limit order book environments: explainability, challenges, and robustness. Proc 4th ACM Int Conf on AI in Finance, p.27-35. https://doi.org/10.1145/3604237.3626854

Das A, Kong W, Sen R, et al., 2024. A decoder-only foundation model for time-series forecasting. Proc 41st Int Conf on Machine Learning, p.10148-10167.

Das V, 2015. Learning Redis. Packt Publishing Ltd., Birmingham, UK.

Daumé HIII, 2007. Frustratingly easy domain adaptation. Proc 45th Annual Meeting of the Association of Computational Linguistics, p.256-263. https://doi.org/10.18653/v1/P18-1183

Dempster M, Kanniainen J, Keane J, et al., 2018. High-Performance Computing in Finance: Problems, Methods, and Solutions (1st Ed.). Chapman and Hall/CRC, New York, NY, USA. https://doi.org/10.1201/9781315372006

Dessaint O, Foucault T, Fresard L, 2024. Does alternative data improve financial forecasting? The horizon effect. *J Fin*, 79(3):2237-2287. https://doi.org/10.1111/jofi.13323

Dixon M, 2022. The book of alternative data: a guide for investors, traders and risk managers. *Quant Fin*, 22(8):1427-1428. https://doi.org/10.1080/14697688.2022.2078736

Dixon W, Yuen K, 1974. Trimming and winsorization: a review. *Stat Hefte*, 15:157-170. https://doi.org/10.1007/BF02922904

Dunning T, Friedman B, Loukides M, et al., 2014. Time Series Databases: New Ways to Store and Access Data. O'Reilly Media, Incorporated.

Engelberg J, McLean RD, Pontiff J, et al., 2023. Do cross-sectional predictors contain systematic information? *J Fin Quant Anal*, 58(3):1172-1201. https://doi.org/10.1017/S0022109022000266

Fabozzi F, Mann S, Fabozzi F, 2021. The Handbook of Fixed Income Securities (9th Ed.). McGraw-Hill Education, New York, USA.

Fama EF, French KR, 1992. The cross-section of expected stock returns. *J Fin*, 47(2):427-465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fang Y, Ren K, Liu W, et al., 2021. Universal trading for order execution with Oracle policy distillation. Proc AAAI Conf on Artificial Intelligence, 35:107-115. https://doi.org/10.1609/aaai.v35i1.16083

Fang Y, Tang Z, Ren K, et al., 2023. Learning multi-agent intention-aware communication for optimal multi-order execution in finance. Proc 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.4003-4012. https://doi.org/10.1145/3580305.3599856

Feng F, He X, Wang X, et al., 2019. Temporal relational ranking for stock prediction. *ACM Trans Inform Syst*, 37(2):1-30. https://doi.org/10.1145/3309547

Feng F, Wang X, He X, et al., 2021. Time horizon-aware modeling of financial texts for stock price prediction. Proc 2nd ACM Int Conf on AI in Finance, p.1-8. https://doi.org/10.1145/3490354.3494416

Gastineau GL, 2001. Exchange-traded funds: an introduction. *J Portf Manag*, 27(3):88-96. https://doi.org/10.3905/jpm.2001.319783

Goodfellow I, Bengio Y, Courville A, 2016. Deep Learning. MIT Press, Cambridge, USA.

Gupta T, Leung E, Roscovan V, 2022. Consumer spending and the cross-section of stock returns. *J Portf Manag*, 48(7):117-137. https://doi.org/10.3905/jpm.2022.1.365

Han J, E H, Le G, et al., 2011. Survey on NoSQL database. 6th Int Conf on Pervasive Computing and Applications, p.363-366. https://doi.org/10.1109/ICPCA.2011.6106531

Han S, Pool J, Tran J, et al., 2015. Learning both weights and connections for efficient neural networks. Proc 29th Int Conf on Neural Information Processing Systems, p.1135-1143.

Hastie T, Tibshirani R, Friedman J, 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed.). Springer, New York, USA.

Ho T, Stoll HR, 1981. Optimal dealer pricing under transactions and return uncertainty. *J Fin Econ*, 9(1):47-73. https://doi.org/10.1016/0304-405X(81)90020-9

Hochreiter S, Schmidhuber J, 1997. Long short-term memory. *Neur Comput*, 9(8):1735-1780.

Howard J, Ruder S, 2018. Universal language model fine-tuning for text classification. Proc 56th Annual Meeting of the Association for Computational Linguistics, p.328-339. https://doi.org/10.18653/v1/P18-1031

Hu EJ, Shen Y, Wallis P, et al., 2021. LoRA: low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685

Hu Z, Liu W, Bian J, et al., 2018. Listening to chaotic whispers: a deep learning framework for news-oriented stock trend prediction. Proc 11th ACM Int Conf on Web Search and Data Mining, p.261-269. https://doi.org/10.1145/3159652.3159690

Hull JC, 2021. Options, Futures and Other Derivatives (11th Ed.). Pearson, Hoboken, NJ, USA.

Jacob B, Kligys S, Chen B, et al., 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2704-2713. https://doi.org/10.1109/CVPR.2018.00286

Jiao X, Li Z, Xu C, et al., 2023. Microstructure-empowered stock factor extraction and utilization. https://doi.org/10.48550/arXiv.2308.08135

Kakushadze Z, 2016. 101 formulaic alphas. *Wilmott*, 2016(84):72-81. https://doi.org/10.1002/wilm.10525

Kim AY, Tse Y, Wald JK, 2016. Time series momentum and volatility scaling. *J Fin Mark*, 30:103-124. https://doi.org/10.1016/j.finmar.2016.05.003

Kou Y, Wang CY, Ye Q, 2013. An empirical study of calendar spread arbitrage based on high-frequency data: the case of CSI 300 index futures. Int Conf on Management Science and Engineering, p.1604-1609. https://doi.org/10.1109/ICMSE.2013.6586481

Krauss C, 2017. Statistical arbitrage pairs trading strategies: review and outlook. *J Econ Surv*, 31(2):513-545. https://doi.org/10.1111/joes.12153

Kwan C, 1999. A note on market-neutral portfolio selection. *J Bank Fin*, 23(5):773-800. https://doi.org/10.1016/S0378-4266(98)00114-9

Lewis P, Perez E, Piktus A, et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33:9459-9474.

Li Y, Wang T, Sun B, et al., 2022. Detecting the lead–lag effect in stock markets: definition, patterns, and investment strategies. *Fin Innov*, 8(1):51. https://doi.org/10.1186/s40854-022-00356-3

Li Y, Huang Y, Ildiz ME, et al., 2024. Mechanics of next token prediction with self-attention. Int Conf on Artificial Intelligence and Statistics, p.685-693.

Liew JKS, Budavari T, Kang Z, et al., 2020. Pairs trading strategy with geolocation data: the battle between Under Armour and Nike. *J Fin Data Sci*, 2(1):126-143. https://doi.org/10.3905/jfds.2019.1.024

Lin S, Beling PA, 2021. An end-to-end optimal trade execution framework based on proximal policy optimization. Proc 29th Int Joint Conf on Artificial Intelligence (Special Track on AI in FinTech), p.4548-4554. https://doi.org/10.24963/ijcai.2020/627

Liu T, Roberts S, Zohren S, 2023. Deep inception networks: a general end-to-end framework for multi-asset quantitative strategies. https://doi.org/10.48550/arXiv.2307.05522

Lo AW, 2010. Hedge Funds: an Analytic Perspective (Revised and Expanded Edition). Princeton University Press, Princeton, NJ, USA. http://www.jstor.org/stable/j.ctt7rq28

Lo AW, Mamaysky H, Wang J, 2000. Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. *J Fin*, 55(4):1705-1765. https://doi.org/10.1111/0022-1082.00265

MacKinlay AC, 1997. Event studies in economics and finance. *J Econ Liter*, 35(1):13-39.

Madhavan A, 2000. Market microstructure: a survey. *J Fin Mark*, 3(3):205-258. https://doi.org/10.1016/S1386-4181(00)00007-0

Mao A, Mohri M, Zhong Y, 2023. Cross-entropy loss functions: theoretical analysis and applications. Proc 40th Int Conf on Machine Learning, p.23803-23828.

Nagy P, Frey S, Sapora S, et al., 2023. Generative AI for end-to-end limit order book modelling: a token-level autoregressive generative model of message flow using a deep state space network. 4th ACM Int Conf on AI in Finance, p.91-99. https://doi.org/10.1145/3604237.3626898

Nie Y, Nguyen NH, Sinthong P, et al., 2023. A time series is worth 64 words: long-term forecasting with Transformers. https://arxiv.org/abs/2211.14730

OpenAI, Achiam J, Adler S, et al., 2024. GPT-4 technical report. https://arxiv.org/abs/2303.08774

Paatela A, Noschis E, Hameri AP, 2017. Abnormal stock returns using supply chain momentum and operational financials. *J Portf Manag*, 43:50-60. https://doi.org/10.3905/jpm.2017.43.2.050

Pan SJ, Yang Q, 2010. A survey on transfer learning. *IEEE Trans Knowl Data Eng*, 22(10):1345-1359. https://doi.org/10.1109/TKDE.2009.191

Păuna C, 2019. Additional limit conditions for breakout trading strategies. *Inform Econ*, 23(2):25-33.

Penman SH, 2013. Financial Statement Analysis and Security Valuation (5th Ed.). McGraw-Hill Education, New York, USA.

Pfeiffer J, Ruckle A, Gurevych I, 2021. AdapterFusion: non-destructive task composition for transfer learning. Proc 16th Conf of the European Chapter of the Association for Computational Linguistics, p.487-503.

Qin M, Sun S, Zhang W, et al., 2024. EarnHFT: efficient hierarchical reinforcement learning for high frequency trading. Proc 38th AAAI Conf on Artificial Intelligence and 36th Conf on Innovative Applications of Artificial Intelligence and 14th Symp on Educational Advances in Artificial Intelligence, p.14669-14676. https://doi.org/10.1609/aaai.v38i13.29384

Salloum S, Dautov R, Chen X, et al., 2016. Big data analytics on Apache Spark. *Int J Data Sci Anal*, 1(3):145-164.

Saputra MYE, Arief SN, Wijayaningrum VN, et al., 2024. Real-time server monitoring and notification system with Prometheus, Grafana, and Telegram integration. ASU Int Conf in Emerging Technologies for Sustainability and Intelligent Systems, p.1808-1813.

Sawhney R, Agarwal S, Wadhwa A, et al., 2020. Spatiotemporal hypergraph convolution network for stock movement forecasting. IEEE Int Conf on Data Mining, p.482-491. https://doi.org/10.1109/ICDM50108.2020.00057

Sawhney R, Agarwal S, Wadhwa A, et al., 2021. Stock selection via spatiotemporal hypergraph attention network: a learning to rank approach. Proc AAAI Conf on Artificial Intelligence, 35(1):497-504. https://doi.org/10.1609/aaai.v35i1.16127

Sharpe WF, 1966. Mutual fund performance. *J Bus*, 39(1):119-138.

Sheikh A, 1996. Barra's risk models. *Barra Res Insights*, 1:1-24.

Shleifer A, Vishny RW, 1997. The limits of arbitrage. *J Fin*, 52(1):35-55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

Shvachko K, Kuang H, Radia S, et al., 2010. The Hadoop distributed file system. IEEE 26th Symp on Mass Storage Systems and Technologies, p.1-10. https://doi.org/10.1109/MSST.2010.5496972

Sorge M, 2004. Stress-testing financial systems: an overview of current methodologies. BIS Working Paper, No. 165. https://doi.org/10.2139/ssrn.759585

Stan CS, Pandelica AE, Zamfir VA, et al., 2019. Apache Spark and Apache Ignite performance analysis. 22nd Int Conf on Control Systems and Computer Science, p.726-733.

Stonebraker M, 2010. SQL databases v. NoSQL databases. *Commun ACM*, 53(4):10-11. https://doi.org/10.1145/1721654.172165

Sun S, Xue W, Wang R, et al., 2022. DeepScalper: a risk-aware reinforcement learning framework to capture fleeting intraday trading opportunities. Proc 31st ACM Int Conf on Information & Knowledge Management, p.1858-1867. https://doi.org/10.1145/3511808.3557283

Sun Y, Liu L, Xu Y, et al., 2024. Alternative data in finance and business: emerging applications and theory analysis. *Fin Innov*, 10(1):127.

Tetlock PC, Saar-Tsechansky M, Macskassy S, 2008. More than words: quantifying language to measure firms' fundamentals. *J Fin*, 63(3):1437-1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

Toth B, Eisler Z, Lillo F, et al., 2012. How does the market react to your order flow? *Quant Fin*, 12(7):1015-1024. https://doi.org/10.1080/14697688.2012.690886

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30:5998-6008.

Wafi AS, Hassan H, Mabrouk A, 2015. Fundamental analysis models in financial markets—review study. *Proc Econ Fin*, 30:939-947. https://doi.org/10.1016/S2212-5671(15)01344-1

Wang J, Zhang Y, Tang K, et al., 2019. AlphaStock: a buying-winners-and-selling-losers investment strategy using interpretable deep reinforcement attention networks. Proc 25th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining, p.1900-1908. https://doi.org/10.1145/3292500.3330647

Wang L, Zhang X, Su H, et al., 2024. A comprehensive survey of continual learning: theory, method and application. *IEEE Trans Patt Anal Mach Intell*, 46(8):5362-5383. https://doi.org/10.1109/TPAMI.2024.3367329

Wang L, Chen S, Jiang L, et al., 2025. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artif Intell Rev*, 58:227. https://doi.org/10.1007/s10462-025-11236-4

Wang S, Yuan H, Zhou L, et al., 2025. Alpha-GPT: human–AI interactive alpha mining for quantitative investment. https://doi.org/10.48550/arXiv.2308.00016

Wang Z, Huang B, Tu S, et al., 2021. DeepTrader: a deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. Proc AAAI Conf on Artificial Intelligence, 35(1):643-650. https://doi.org/10.1609/aaai.v35i1.16144

Wei Z, Dai B, Lin D, 2023. E2EAI: end-to-end deep learning framework for active investing. Proc 4th ACM Int Conf on AI in Finance, p.55-63. https://doi.org/10.1145/3604237.3626848

Xia SQ, Simonian J, 2021. Measuring investment skill in multi-asset strategies: an empirical study of the information coefficient as weighted rank correlation. *J Portf Manag*, 47(4):135-144. https://doi.org/10.3905/jpm.2021.1.208

Xie X, Liu H, Hou W, et al., 2023. A brief survey of vector databases. 9th Int Conf on Big Data and Information Analytics, p.364-371.

Xu Y, Cohen SB, 2018. Stock movement prediction from Tweets and historical prices. Proc 56th Annual Meeting of the Association for Computational Linguistics, p.1970-1979. https://doi.org/10.18653/v1/P18-1183

Yang W, Wei Y, Wei H, et al., 2023. Survey on explainable AI: from approaches, limitations and applications aspects. *Human-Centr Intell Syst*, 3:161-188. https://doi.org/10.1007/s44230-023-00038-y

Yin S, Fu C, Zhao S, et al., 2024. A survey on multimodal large language models. *Nat Sci Rev*, 11(12). https://doi.org/10.1093/nsr/nwae403

Yosinski J, Clune J, Bengio Y, et al., 2014. How transferable are features in deep neural networks? Proc 28th Int Conf on Neural Information Processing Systems, p.3320-3328.

Yu H, Hao X, Wu L, et al., 2023. Eye in outer space: satellite imageries of container ports can predict world stock returns. *Human Soc Sci Commun*, 10(1):1-16. https://doi.org/10.1057/s41599-023-01891-9

Zhang F, Guo R, Cao H, 2020. Information coefficient as a performance measure of stock selection models. https://arxiv.org/abs/2010.08601

Zhang L, Aggarwal C, Qi GJ, 2017. Stock price prediction via discovering multi-frequency trading patterns. Proc 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, p.2141-2149. https://doi.org/10.1145/3097983.3098117

Zhang Y, Zhao P, Wu Q, et al., 2022. Cost-sensitive portfolio selection via deep reinforcement learning. *IEEE Trans Knowl Data Eng*, 34(1):236-248. https://doi.org/10.1109/TKDE.2020.2979700

Zhang Z, Zohren S, Roberts S, 2019. DeepLOB: deep convolutional neural networks for limit order books. *IEEE Trans Signal Process*, 67(11):3001-3012. https://doi.org/10.1109/TSP.2019.2907260

Zhang Z, Lim B, Zohren S, 2021. Deep learning for market by order data. *Appl Math Fin*, 28(1):79-95. https://doi.org/10.1080/1350486X.2021.1967767

Zhao J, Zhang C, Qin M, et al., 2025. QuantFactor REINFORCE: mining steady formulaic alpha factors with variance-bounded REINFORCE. *IEEE Trans Signal Process*, 73:2448-2463. https://doi.org/10.1109/TSP.2025.3576781

Zhong L, Wu J, Li Q, et al., 2023. A comprehensive survey on automatic knowledge graph construction. *ACM Comput Surv*, 56(4):1-62. https://doi.org/10.1145/3618295

Zhou C, Li Q, Li C, et al., 2024. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. *Int J Mach Learn Cybern*. https://doi.org/10.1007/s13042-024-02443-6