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

MENTOR: a multi-agent framework for event and narrative trend prediction with optimized reasoning#

Liyuan CHEN[1,2], Gaoguo JIA[2], Dongsheng GU[2], Jiangpeng YAN[2], Yuhang JIANG[2], Xiu LI[1], Xiaojun ZENG†‡[3]

[1]*Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China*

[2]*E Fund Management Co., Ltd., Guangzhou 510620, China*

[3]*Department of Computer Science, The University of Manchester, Manchester M139PL, UK*

†E-mail: x.zeng@manchester.ac.uk

Received Aug. 30, 2025; Revision accepted Oct. 13, 2025; Crosschecked Oct. 21, 2025

Abstract: Narrative economics suggests that financial markets are strongly influenced by evolving narratives, creating opportunities for forecasting emerging events and their economic impacts. However, existing large language model (LLM)-based approaches are inadequate in terms of systematic task decomposition and alignment with financial applications. We propose MENTOR, a multi-agent framework for event and narrative trend prediction that integrates teacher–student iterative reasoning with progressive subtasks: detecting and ranking trending events, forecasting future events from current narratives, and predicting industry index performance influenced by these events. Experiments on our self-constructed Chinese key opinion leader (KOL) articles dataset and English financial news dataset show that MENTOR consistently outperforms recent baselines such as the stakeholder-enhanced future event prediction (StkFEP) and summarize–explain–predict (SEP) frameworks in both event prediction and industry ranking tasks. In addition, the backtest results at the portfolio level show that improved event and industry forecasts can bring about a practical improvement in investment performance. These results demonstrate that incorporating structured reasoning and multi-agent feedback enables more reliable event forecasting and strengthens the connection between narrative dynamics and financial market outcomes.

Key words: Narrative economics; Multi-agent; Event detection; Event forecasting
https://doi.org/10.1631/FITEE.2500608 **CLC number:** TP18; F830.9

1 Introduction

Narrative economics emphasizes that financial markets are driven not solely by fundamentals but also by the narratives that circulate among investors and the public (Shiller, 2019). Narratives shape expectations, guide sentiment, and can amplify or mitigate economic fluctuations. In practice, analysts often anticipate emerging narratives and then map them to likely sectoral or market responses. This decomposition, predicting events first and then inferring financial impacts, offers a structured way to avoid modeling the randomness inherent in price movements. However, automating this reasoning process remains an open challenge.

Although narratives can spread in complex and sometimes unpredictable ways, Robert Shiller's theory of narrative economics provides a foundation for structured forecasting (Shiller, 2019). Specifically, Shiller argues that (1) new narratives typically emerge from existing ones through association or amplification, (2) economically relevant narratives tend to recur across different historical episodes

‡ Corresponding author

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2500608) contains supplementary materials, which are available to authorized authors

ORCID: Liyuan CHEN, https://orcid.org/0009-0005-9710-9719; Xiaojun ZENG, https://orcid.org/0000-0002-2320-2495

(e.g., inflation fears and technological disruption), and (3) despite surface-level diversity, narratives can be meaningfully clustered by theme and market impact. These properties imply a degree of bounded predictability: although we cannot foresee every narrative shock, we can identify high-potential trajectories rooted in current discourse. As illustrated in Fig. 1, our framework operationalizes this insight by analyzing the narrative constellation of today to forecast future events and their financial implications.

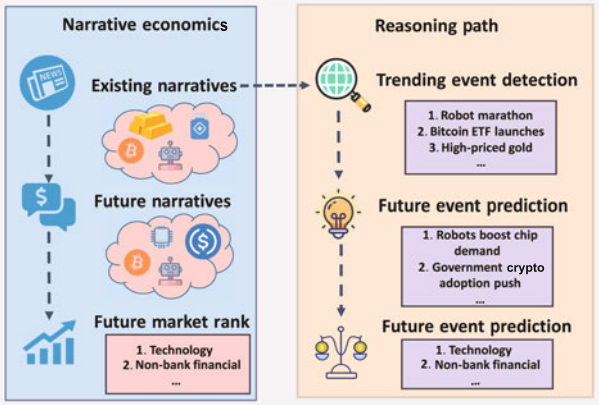

Fig. 1 Narrative economics: theoretical framework and analytical reasoning path. This figure illustrates our motivation, grounded in the theory that markets are narrative-driven. New narratives evolve from existing ones, propelling market movements. Our proposed framework analyzes current narrative trends to predict future events and their narratives, providing a reliable bridge for forecasting. This approach allows analysts to navigate complex market information and noisy price data with greater clarity

Existing event prediction approaches often treat future event forecasting as either a classification task or a generative task, relying on manually curated datasets that quickly become outdated. While large language models (LLMs) provide powerful contextual reasoning ability, they struggle in open-ended prediction settings and frequently lack alignment with financial applications. Conversely, LLM-based financial agents (Yang HY et al., 2023; Yu YY et al., 2023, 2024) focus on stock-level decisions such as buy/sell signals. These systems generally simplify reasoning into binary or single-asset predictions, which limits their applicability to multi-asset financial forecasting and prevents a principled connection to narrative-driven events.

To address these gaps, we propose MENTOR, a multi-agent framework for event and narrative trend prediction that explicitly couples event forecasting with financial index ranking. MENTOR decomposes the problem into three progressive tasks: (1) detecting and ranking trending events, (2) forecasting future events that are likely to emerge from current narratives, and (3) predicting and ranking industry indices based on these events. A teacher–student iterative reasoning mechanism further refines prediction quality: teacher agents provide feedback to student agents across both event-level and industry-level tasks, enabling systematic improvement of reasoning strategies. By bridging narrative event prediction with sectoral financial forecasting, MENTOR provides a pathway for operationalizing narrative economics within large-scale data-driven settings.

Our experiments on datasets from Chinese key opinion leaders (KOL) and English financial news show that MENTOR successfully predicts approximately 50% of future hotspot events, with an average prediction score exceeding 4.5 out of 10. This provides a more rigorous and quantifiable evaluation framework for the task of future event prediction. Additionally, MENTOR surpasses state-of-the-art models in ranking the returns of 11 S&P industries and 9 A-share industries, effectively solving the multi-asset ranking problem that is challenging for prior financial agent frameworks. Our contributions are summarized as follows:

1. We propose a novel multi-agent framework, MENTOR, which emulates human analysts by decomposing complex prediction tasks into manageable subtasks and iteratively enhancing reasoning through teacher–student collaborations.

2. We demonstrate that MENTOR effectively predicts future narrative trends, significantly improving the precision of trending event predictions.

3. We provide empirical evidence of MENTOR's effectiveness across diverse datasets, showcasing its capability in multi-asset ranking tasks and its potential for practical applications in financial forecasting and narrative trend analysis.

2 Related works

2.1 Long-text reasoning for trending events

In the field of big data analytics for financial news, effective trending events are defined as occurrences that generate high engagement on social

media platforms and exhibit considerable persistence over time. Identifying effective trending events presents a significant challenge.

In the expanding field of LLM applications, the ability to process long contexts has become increasingly important (Zheng et al., 2023; Gao et al., 2024). Conventional Transformer-based models face significant computational challenges when scaling to extensive text due to the quadratic complexity of their self-attention mechanisms. Developing models capable of handling longer contexts is an active area of research (Wang et al., 2024). However, despite advancements in computational efficiency, recent findings suggest that long-context models remain less effective in utilizing their broader context (Liu NF et al., 2024). The parallel context window (PCW) (Ratner et al., 2023) is a method that segments long contexts into multiple blocks and restricts the attention mechanism to operate within each window. Another innovative approach is StreamingLLM (Xiao et al., 2024), which addresses the "attention sink" phenomenon. This phenomenon occurs when a majority of attention scores are allocated to the initial tokens, regardless of their relevance. Additionally, MemGPT (Packer et al., 2024) aims to overcome the limitations of fixed-length context windows in traditional LLMs.

2.2 Future event prediction

Future event prediction (FEP) focuses on forecasting potential future outcomes based on historical events and precursors (Zhao, 2022). However, most of these works frame event prediction as a classification problem, limiting the outcomes to a predefined set of categories or labels. For example, predictions are constrained to binary yes/no decisions (Ye et al., 2024) or to a fixed set of predefined candidates (Zhu FQ et al., 2023).

Despite these classification-oriented approaches, time-series forecasting models are also leveraged to predict continuous market indicators and risk metrics as part of FEP. Classical statistical models such as ARIMA (Shumway and Stoffer, 2019) have been adapted to forecast stock price movements beyond categorical labels, demonstrating high baseline performance (Ariyo et al., 2014). More advanced recurrent neural networks, including long short-term memory (LSTM) and gated recurrent unit (GRU) architectures, capture nonlinear dependencies and temporal dynamics inherent in financial time series (Mroua and Lamine, 2023; Zhu P et al., 2024; Bao et al., 2025). These continuous-output models transcend fixed label sets, enabling more granular and actionable future event predictions in financial markets. Orthogonal to these temporal approaches, another prominent line of research leverages graph-based models to explicitly capture structural dependencies among financial entities. Such methods range from employing news-driven techniques for portfolio selection (Liang et al., 2021; Yang MY et al., 2022) to constructing complex relational graphs from financial texts and knowledge bases for stock movement prediction (Liu MP et al., 2024; Qian et al., 2024; Shi et al., 2024).

Due to the remarkable performance of LLMs (Llama Team, 2024), some studies leverage their powerful contextual understanding abilities in the FEP task (Li ML et al., 2021). Nevertheless, it remains a challenging task for LLMs due to the demand for advanced reasoning and prospection capabilities in open-ended FEP. OpenEP (Guan et al., 2024) constructs the OpenEPBench dataset and proposes the stakeholder-enhanced future event prediction (Stk-FEP) framework that incorporates stakeholders to enhance prediction accuracy. The experiments suggest that although powerful in text generation, LLMs may fall short in providing accurate and detailed predictions without specific guidance or enhancements.

2.3 Agent systems for financial decision-making

Language agent systems designed for financial decision-making, such as FinGPT (Yang HY et al., 2023), FinMem (Yu YY et al., 2023), and FinAgent (Zhang WT et al., 2024), have demonstrated high performance in financial tasks, leveraging LLMs to process financial data and assisting in decision-making processes. These studies attempt to model the impact of narratives on market behavior using agent-based computational economics (ACE) and opinion dynamics. For instance, SEP (Koa et al., 2024) uses self-reflective agents and proximal policy optimization (PPO) to train LLMs to generate explainable stock predictions, generalizing well to portfolio construction tasks. FinCon (Yu YY et al., 2024) features a manager–analyst hierarchy and a dual-level risk control mechanism, achieving superior performance through conceptual verbal

reinforcement and optimized information synthesis. These methods integrate opinion dynamics into trading behaviors, allowing for simulations where agent's opinions can drive prices, and prices can, in turn, alter opinions. This bidirectional influence aligns with Shiller's notion of narrative economics.

However, existing language agent systems are often limited to single-asset trading tasks, reducing their adaptability to multi-asset financial applications like portfolio management. Financial markets are inherently complex and involve a diverse range of assets; therefore, a system that can handle multi-asset scenarios is more practical and valuable for real-world applications. Our proposed framework, MENTOR, addresses these limitations by introducing a multi-agent event and narrative trend prediction system augmented with optimized reasoning. MENTOR emphasizes the mutual influence between narratives and financial activities.

3 Problem formulation

The MENTOR framework is proposed to answer three questions.

3.1 Q1: How can we detect trending events from a large amount of real-time financial news?

For each time interval, the complete textual information $\mathcal{N}$ (comprising financial news or KOL articles) contains several trending events. These trending events can be ranked based on heat-related evaluation metrics. We optimize the following method to ensure that the obtained ranking $\mathcal{H}$ closely approximates the future heat ranking of these events, formulated as

$$\mathcal{H} = \delta(\mathcal{N}), \tag{1}$$

where δ represents the extraction and ranking operator.

3.2 Q2: How can we predict future events based on current hot topics?

According to narrative economics, future trending events emerge from existing trending events. Therefore, we attempt to further predict and rank the trending events that will occur in the next time interval based on the ranking of trending events obtained for Q1. Unlike Q1, Q2 requires predicting events that have not yet occurred, formulated as

$$\hat{\mathcal{H}} = \gamma(\mathcal{H}), \tag{2}$$

where γ is the future event prediction operator.

3.3 Q3: Can we forecast industry index returns using detected and predicted events?

After obtaining the predictions of future events for Q2, we can predict the performance rankings of industry indices in the next time interval based on these predicted events and the performance of industry indices during the current time interval, formulated as

$$\hat{\boldsymbol{R}} = \omega\left(\mathcal{H}, \hat{\mathcal{H}}, \boldsymbol{R}\right), \tag{3}$$

where ω is the industry ranking operator.

4 Methodology

4.1 Framework overview

We propose MENTOR, a multi-agent framework that operationalizes narrative economics by decomposing the forecasting pipeline into three progressive stages:

1. Event detection. Narratives are first extracted from heterogeneous text streams (financial news, social media posts, and KOL commentaries). A clustering procedure groups semantically related narratives into coherent events, which are then ranked by relevance.

2. Future event prediction. Building on these detected events, the system forecasts plausible emerging events that may appear in subsequent time windows. This task requires extrapolating narrative trajectories rather than simply classifying observed events.

3. Industry index ranking. Finally, the framework maps predicted events to sector-level financial indices, producing a ranked list of industries expected to benefit or suffer from the anticipated narratives.

Unlike one-shot prompting, MENTOR employs student agents that engage in the prediction task and teacher agents that provide structured feedback. This design introduces an iterative reasoning mechanism that progressively improves prediction quality across subtasks.

4.2 Agent for trending event detection

4.2.1 Baseline for trending event detection

Given the unpredictability and dynamism of financial news and social media data streams, recent studies have increasingly integrated deep neural networks with unsupervised clustering methods (Li QZ et al., 2023). Specifically, these studies utilize pre-trained embeddings to vectorize text. The embeddings are then employed for clustering to facilitate incremental event detection. Building upon this methodological framework, we first establish a baseline workflow for event detection. Specifically, we adopt the Qwen3-Embedding-8B model, a state-of-the-art embedding model trained on large-scale data (Enevoldsen et al., 2025; Zhang YZ et al., 2025), which currently demonstrates superior performance in multilingual clustering tasks, and combine it with an optimal density-based clustering algorithm (Schubert et al., 2017). This integration constitutes the baseline model for comparative analysis in our study.

4.2.2 Long-text reasoning for trending event detection

We develop a workflow for detecting trending events in long texts, which consists of two phases: large model inference for the extraction of hot events, and retrieval and ranking of these events.

In the extraction phase, we investigate how to simulate an unlimited contextual environment using a fixed context model. Our method draws inspiration from virtual memory paging, designed to allow applications to manage datasets exceeding available memory by transferring data between main memory and disk (Packer et al., 2024). The workflow, as depicted in Fig. 2, consists of several modules. During the knowledge management phase, we effectively combine short-term dialogue memory and long-term historical knowledge through functional executors to manage the infinite flow of long text input. At the same time, in each round of dialogue, we perform specific tasks using a limited context window. This includes system instructions for storing core task prompts, main context for handling event detection and the core tasks of knowledge iteration, and a message queue for storing short-term historical rolling dialogues to capture recent information.

In the retrieval and ranking phase, we use the outputs obtained above as a candidate list of trending events, which also undergoes state-of-the-art vector recall, coarse ranking, and fine ranking processes to accurately match similar trending events in the original news and form aggregated event clusters (Chen et al., 2024; Zhang YZ et al., 2025). We transform the unsupervised clustering task into a supervised retrieval problem, enhancing the accuracy of event detection. Additionally, we define advanced evaluation metrics to assist in the retrospective validation of event effectiveness over time.

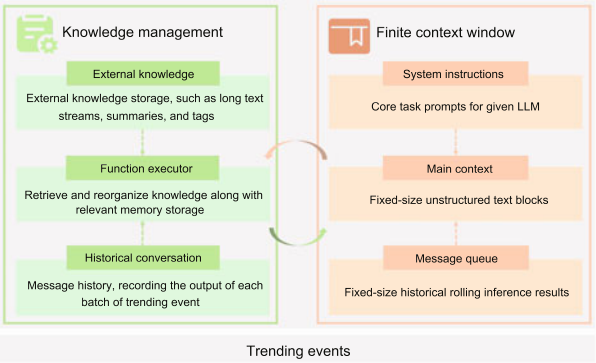

Fig. 2 Workflow of the event detection agent

4.3 Prelude to iterative prediction

After obtaining the trending event ranking via the trending event detection agent (Section 4.2), we further explore how to predict unoccurring hot events, which is the basis for subsequent iterative prediction. By utilizing the trending event detection agent, we have identified which hot events that have already occurred may gain higher levels of discussion in the future. However, since people are generally aware of these events and the market has often already responded, predicting potential hot events is both more valuable and more challenging.

According to theories in narrative economics, narratives form narrative constellations, from which new narratives continually emerge. Consequently, newly occurring hot events are also predictable. Although sudden events are inherently difficult to foresee, the vast majority of hot events can be extrapolated from prior ones. For example, discussions often increase before a company releases its financial reports, and small-scale conflicts can significantly impact energy and gold markets.

Nevertheless, performing such reasoning is highly challenging due to the integration of

multi-dimensional information. Reasoning-focused models like O1 demonstrate strong inferential capabilities when explaining past events, but when it comes to predicting future events, they often overemphasize individual events and tend to provide mediocre answers lacking incremental value. In contrast, generation-focused models have lower computational costs and are more adaptable in predicting future events, but their lack of reasoning ability can lead to illogical predictions.

To address this challenge, we have developed a multi-teacher, multi-student model that iteratively employs teacher agents and reasoning agents to instruct students in enhanced reasoning.

4.4 Iterative reasoning process

4.4.1 Text-based strategy optimization via multi-agent feedback

We conduct rolling backward training along the time axis on a weekly basis, specifically divided into two tasks: event prediction and industry ranking. Each task features a cyclic iterative system, composed of a student agent, a teacher agent, and a reasoning agent (Fig. 3).

We adopt textual gradient descent (TextGrad) (Yuksekgonul et al., 2024) to update the strategies for the two tasks: event prediction and industry ranking. TextGrad is a recently proposed optimization paradigm that improves AI systems by iteratively backpropagating text-based feedback, enabling gradient-like updates without relying on access to model parameters. This approach offers two primary advantages: First, it reduces dependence on carefully engineered initial prompts, allowing the system to start from relatively simple prompts and thereby mitigating the potential overfitting risks common in most prior multi-agent frameworks based on prompt engineering. Second, it enables dynamic adaptation to evolving markets, preventing rigid investment analysis frameworks from failing in the face of market fluctuations.

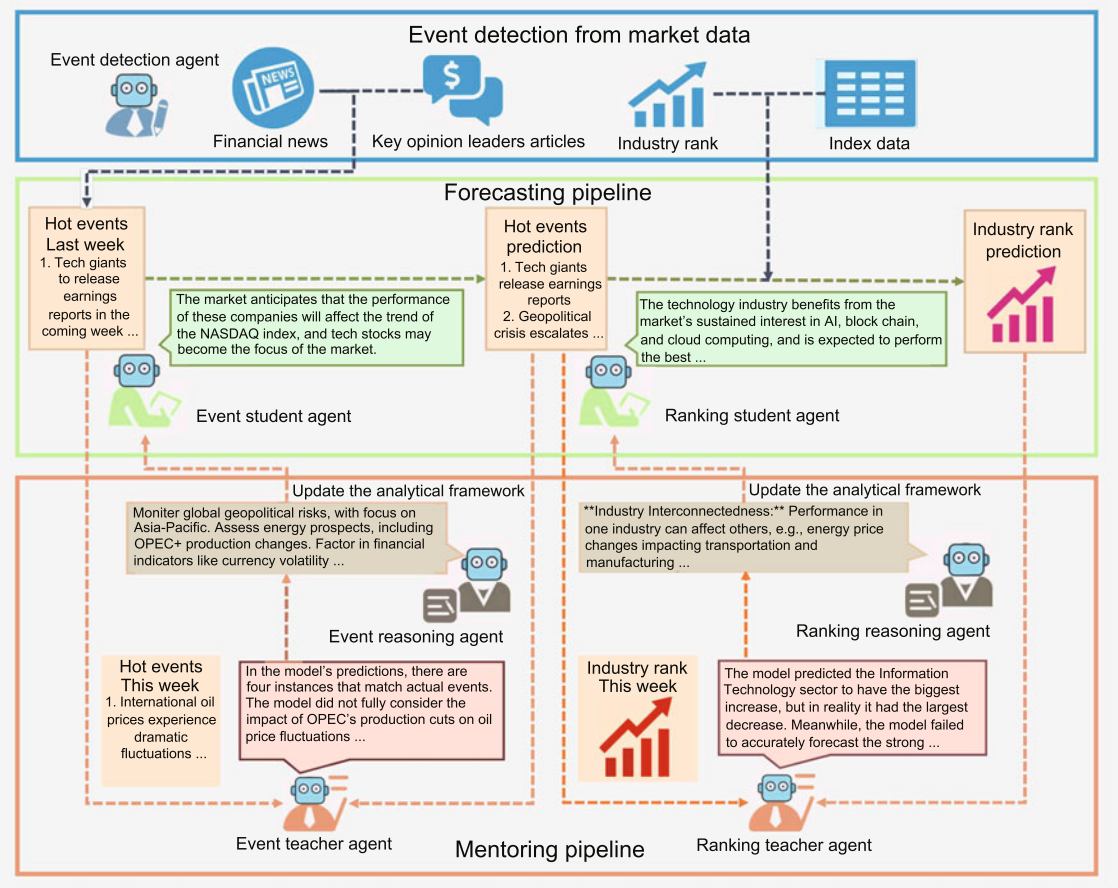

Fig. 3 Illustration of the iterative process in the MENTOR framework

To improve reproducibility and facilitate community adoption, we briefly summarize the design principles underlying the prompt templates used by both teacher and student agents. These prompts are engineered according to two key principles:

1. Conciseness. Aligned with the TextGrad optimization paradigm, prompts are intentionally minimal, providing only essential task instructions without over-constraining reasoning pathways. This encourages the LLM to autonomously construct contextually grounded logic while remaining focused on the prediction objective.

2. Adaptability. Recognizing that different LLM backbones (e.g., DeepSeek and O1) may exhibit sensitivity to phrasing due to safety filters or internal formatting policies, we implement a multi-template fallback mechanism. If a primary prompt fails to elicit a valid or coherent response, the system dynamically selects an alternative formulation with equivalent semantic intent.

Together, these principles support a reproducible and model-agnostic prompting strategy that balances flexibility with task fidelity. The details of prompt initialization and update templates are provided in the supplementary materials. Building on the design principles, the student strategies are formalized and iteratively refined through the training process described below.

1. Initialization

At iteration $n=0$, the student strategies for the two tasks are initialized:

$\mathcal{S}_{\text{event}}^{(0)}$: initial reasoning strategy for event prediction.

$\mathcal{S}_{\text{ranking}}^{(0)}$: initial reasoning strategy for industry ranking.

2. Event prediction

At iteration n, the event strategy predicts the set of likely future events:

$$\hat{\mathcal{H}}^{(n)}=\mathcal{S}_{\text{event}}^{(n-1)}\big(\delta(\mathcal{N}^{(n-1)})\big), \tag{4}$$

where δ extracts salient events from the textual data $\mathcal{N}^{(n-1)}$. This step is analogous to the forward pass in TextGrad, producing predictions based on the current strategy.

3. Industry ranking prediction

The ranking strategy then predicts industry rankings conditioned on the predicted events and past rankings:

$$\hat{\boldsymbol{R}}^{(n)}=\mathcal{S}_{\text{ranking}}^{(n-1)}\big(\hat{\mathcal{H}}^{(n)},\boldsymbol{R}^{(n-1)}\big). \tag{5}$$

This corresponds to another forward pass where the predicted events inform downstream reasoning, akin to multi-task TextGrad propagation.

4. Feedback from teacher models

Teacher agents $\mathcal{T}_{\text{event}}$ and $\mathcal{T}_{\text{ranking}}$ generate feedback signals:

$$\begin{cases}\Delta\mathcal{S}_{\text{ranking}}^{(n)}=\mathcal{T}_{\text{ranking}}\big(\hat{\mathcal{H}}^{(n)},\mathcal{H}^{(n)},\hat{\boldsymbol{R}}^{(n)},\boldsymbol{R}^{(n)},\mathcal{S}_{\text{ranking}}^{(n-1)}\big),\\ \Delta\mathcal{S}_{\text{event}}^{(n)}=\mathcal{T}_{\text{event}}\big(\hat{\mathcal{H}}^{(n)},\mathcal{H}^{(n)},\mathcal{S}_{\text{event}}^{(n-1)},\Delta\mathcal{S}_{\text{ranking}}^{(n)}\big),\end{cases} \tag{6}$$

which play the role of textual gradients in TextGrad, providing corrective signals for the strategies.

5. Strategy update via the gradient-like operator

The strategies are then updated using gradient-like operators $\mathcal{G}_{\text{event}}$ and $\mathcal{G}_{\text{ranking}}$:

$$\begin{cases}\mathcal{S}_{\text{event}}^{(n)}=\mathcal{G}_{\text{event}}\big(\mathcal{S}_{\text{event}}^{(n-1)},\Delta\mathcal{S}_{\text{event}}^{(n)}\big),\\ \mathcal{S}_{\text{ranking}}^{(n)}=\mathcal{G}_{\text{ranking}}\big(\mathcal{S}_{\text{ranking}}^{(n-1)},\Delta\mathcal{S}_{\text{ranking}}^{(n)}\big),\end{cases} \tag{7}$$

highlighting the analogy to gradient descent in TextGrad where textual feedback iteratively refines the model.

The above represents a complete training process, which is iteratively conducted multiple times through rolling training. Once the teacher model has completed the reasoning instruction, the student model can be directly used for reasoning, enabling fast and cost-effective prediction and ranking. Relevant studies on TextGrad (Yuksekgonul et al., 2024) have shown that this method can maintain the results of various tasks at a stable value. Therefore, we can consider that the model iterations will reach a relatively optimal stable state, which has also been verified by experiments.

4.4.2 A concrete example of TextGrad in action

We illustrate one iteration of Algorithm 1 with a realistic weekly forecasting episode. The sequence of events and agent behaviors is as follows:

Input at week $n-1$: The news stream $\mathcal{N}^{(n-1)}$ includes: (1) "Major technology firms will release earnings next week;" (2) "A geopolitical conflict erupts in a key oil-producing region;" (3) "OPEC announces a reduction in oil production."

Algorithm 1 MENTOR: rolling strategy evolution

Require: weekly news $\{\mathcal{N}^{(n)}\}$ and weekly returns $\{\boldsymbol{R}^{(n)}\}$ for $n = 1, 2, \ldots, N$
Ensure: predictions $\hat{\mathcal{H}}^{(n)}, \hat{\boldsymbol{R}}^{(n)}$ for each weekly time step n
$\mathcal{S}_{\text{event}} \leftarrow \mathcal{S}_{\text{event}}^{(0)}$
$\mathcal{S}_{\text{ranking}} \leftarrow \mathcal{S}_{\text{ranking}}^{(0)}$
for each time step $n = 1, 2, \ldots, N$ **do**
  $\mathcal{H}_{\text{last}} \leftarrow \delta(\mathcal{N}^{(n-1)})$
  $\boldsymbol{R}_{\text{last}} \leftarrow \boldsymbol{R}^{(n-1)}$
  $\hat{\mathcal{H}}^{(n)} \leftarrow \mathcal{S}_{\text{event}}(\mathcal{H}_{\text{last}})$
  $\hat{\boldsymbol{R}}^{(n)} \leftarrow \mathcal{S}_{\text{ranking}}(\hat{\mathcal{H}}^{(n)}, \boldsymbol{R}_{\text{last}})$
  Observe ground truth $\mathcal{H}^{(n)}$, $\boldsymbol{R}^{(n)}$
  $\Delta\mathcal{S}_{\text{ranking}} \leftarrow \mathcal{T}_{\text{ranking}}(\hat{\mathcal{H}}^{(n)}, \mathcal{H}^{(n)}, \hat{\boldsymbol{R}}^{(n)}, \boldsymbol{R}^{(n)}, \mathcal{S}_{\text{ranking}})$
  $\Delta\mathcal{S}_{\text{event}} \leftarrow \mathcal{T}_{\text{event}}(\hat{\mathcal{H}}^{(n)}, \mathcal{H}^{(n)}, \mathcal{S}_{\text{event}}, \Delta\mathcal{S}_{\text{ranking}})$
  $\mathcal{S}_{\text{event}} \leftarrow \mathcal{G}_{\text{event}}(\mathcal{S}_{\text{event}}, \Delta\mathcal{S}_{\text{event}})$
  $\mathcal{S}_{\text{ranking}} \leftarrow \mathcal{G}_{\text{ranking}}(\mathcal{S}_{\text{ranking}}, \Delta\mathcal{S}_{\text{ranking}})$
  Output $(\hat{\mathcal{H}}^{(n)}, \hat{\boldsymbol{R}}^{(n)})$
end for

Prediction by $\mathcal{S}_{\text{event}}$: Based on its initial reasoning template, it predicts the upcoming event narrative $\hat{\mathcal{H}}^{(n)}$ as "technology earnings announcements" and "geopolitical escalation."

Prediction by $\mathcal{S}_{\text{ranking}}$: Using $\hat{\mathcal{H}}^{(n)}$ and last week's returns $\boldsymbol{R}^{(n-1)}$, it outputs a predicted ranking vector $\hat{\boldsymbol{R}}^{(n)}$ that assigns the highest score to the technology sector.

Observed outcome in week n: The true event narrative $\mathcal{H}^{(n)}$ is dominated by oil-market dynamics, and the true return vector $\boldsymbol{R}^{(n)}$ shows high outperformance in the energy sector and notable underperformance in technology.

Feedback from $\mathcal{T}_{\text{event}}$: The teacher agent of the event prediction task compares $\hat{\mathcal{H}}^{(n)}$ against $\mathcal{H}^{(n)}$, and provides: "Treat coordinated supply shocks (e.g., OPEC cuts) and conflicts in energy-producing regions as high-priority triggers even when scheduled corporate events exist."

Feedback from $\mathcal{T}_{\text{ranking}}$: The teacher agent of the industry ranking task compares $\hat{\boldsymbol{R}}^{(n)}$ against $\boldsymbol{R}^{(n)}$, and provides: "When $\mathcal{S}_{\text{event}}$ flags a commodity-driven event, increase sensitivity to energy-related sectors and reduce exposure to input-cost-sensitive sectors like technology."

Update by $\mathcal{G}_{\text{event}}$: The event reasoning agent adds a new conditional rule to $\mathcal{S}_{\text{event}}$: "If OPEC announces production cuts and a conflict occurs in an oil-producing region, prioritize commodity-driven narratives over scheduled corporate events."

Update by $\mathcal{G}_{\text{ranking}}$: The ranking reasoning agent refines the ranking logic in $\mathcal{S}_{\text{ranking}}$ to include: "When $\mathcal{S}_{\text{event}}$ assigns high confidence to oil-related events, downgrade technology and upgrade energy in $\hat{\boldsymbol{R}}^{(n)}$."

Result in subsequent weeks: When similar signals appear, $\mathcal{S}_{\text{event}}$ correctly predicts $\hat{\mathcal{H}}^{(n)}$ centered on oil volatility, and $\mathcal{S}_{\text{ranking}}$ produces a more accurate $\hat{\boldsymbol{R}}^{(n)}$ demonstrating adaptive, interpretable learning.

This example shows that TextGrad's updates are not black-box parameter adjustments, but structured, human-readable edits to reasoning logic, mediated by reasoning agents and grounded in diagnostic feedback.

5 Experiments

5.1 Datasets

To comprehensively evaluate MENTOR, we use two datasets:

1. Chinese KOL articles dataset: This dataset comprises posts and interactions from influential figures on Chinese social media platforms, reflecting trending narratives and public sentiments within the Chinese market. A total of 160 190 posts were collected from these influential figures, commonly referred to as "Big Vs," during the period from January 1, 2023 to December 31, 2024.

2. English financial news dataset: This dataset comprises financial news articles and reports sourced from internationally recognized English-language media outlets, including Yahoo News, Investing.com, Street Insider, and MENAFN. It captures global events and their potential impact on financial markets, providing a comprehensive resource for analysis. A total of 72 108 raw financial news articles and reports were collected during the same period, aligning with the timeframe of the Chinese KOL articles dataset.

For detailed information about the two datasets mentioned above, please refer to the supplementary materials. For industry classification, price fluctuations, and ranking data, our S&P data are sourced from Bloomberg, while A-share data are obtained from Wind Information.

5.2 Evaluation metrics

5.2.1 Evaluation for trending event detection

Through the aforementioned retrieval and ranking process for hot events, we can derive evaluation

metrics for event detection across four dimensions: event repetition rate (ERR), influencer coverage rate (ICR), key point quantity (KPQ), and sentiment disagreement index (SDI). Detailed definitions and calculation methods of these four metrics are described in the supplementary materials. Furthermore, based on the dynamic evolution mechanism of public opinion events, entropy serves as a valuable metric for identifying critical time nodes during their evolution (Eimann, 2008; Ma et al., 2014). By analyzing entropy variations, the evolutionary stage of a public opinion event can be determined, enabling the assessment of its temporal effectiveness. In this study, we validate event effectiveness by defining entropy increase metrics across various time intervals and utilize the calculated accuracy to evaluate the model's performance.

5.2.2 Evaluation for trending event prediction

For the event prediction task, we evaluate performance by measuring the degree of overlap between the predicted events and the actual events. For the ranking task, we utilize the Spearman rank correlation coefficient and the Kendall τ rank correlation coefficient, and assess the overlap of the top n industries.

Event prediction score:

$$\text{Score}^{(n)} = \left|\hat{\mathcal{H}}^{(n)} \cap \mathcal{H}^{(n)}\right|. \tag{8}$$

Spearman's rank correlation coefficient:

$$\rho^{(n)} = 1 - \frac{6\sum_{i=1}^{K}\left(r_i^{(n)} - s_i^{(n)}\right)^2}{K(K^2-1)}. \tag{9}$$

Kendall's τ coefficient:

$$\tau^{(n)} = \frac{2\left(C^{(n)} - D^{(n)}\right)}{K(K-1)}. \tag{10}$$

Herein, K is the number of industries, $r_i^{(n)}$ is the rank of industry i in the predicted rankings $\hat{\boldsymbol{R}}^{(n)}$, $s_i^{(n)}$ is the rank of industry i in the actual rankings $\boldsymbol{R}^{(n)}$, and $C^{(n)}$ and $D^{(n)}$ are the numbers of concordant and discordant pairs, respectively.

5.3 Baselines

Since the MENTOR framework encompasses multiple sub-tasks, we compare each task with its corresponding optimal algorithm. We will further discuss the contribution of each component to the overall performance in Section 5.5.

5.3.1 Trending event detection

In the event detection phase, we first employ the workflow outlined in Section 4.2, including an optimal embedding model and a density-based clustering algorithm, as the baseline for comparison. Furthermore, we utilize DeepSeek and O1 as foundational models for trending event detection, thereby enhancing the capability of extracting hotspots through the incorporation of a short-term memory augmentation module. Finally, we develop a long-text reasoning agent for trending event detection by integrating long- and short-term memory with dialogue workflows.

5.3.2 Event prediction

Regarding event prediction, the vast number of news articles far exceeds the token limits that LLMs can process. Therefore, we compare our approach with the current leading future event prediction framework, StkFEP. Although the final output domain is not exactly the same as that in the original paper, we adapt StkFEP to accomplish this task because it supports open-ended responses. Considering the emergence of more advanced baseline models, we separately test StkFEP using DeepSeek and O1 as baseline models.

5.3.3 Industry indices ranking

For the industry indices ranking task, although there is substantial research related to multi-agent portfolio management, their applications are often limited to a small number of individual stocks. The volume of text data used and the number of assets that need to be ranked are both significantly smaller than the counterparts in our current task. Therefore, we have selected the SEP framework, which can be adapted to this task, as a reference. Since the inputs are all purely textual, we need only to convert the final asset allocation proportions in the portfolio into asset rankings. In addition, we have conducted direct predictions using DeepSeek and O1. To ensure that their input token lengths meet the requirements, the textual inputs for all comparative models rely on the trending events from the first step. Furthermore, we

have included Momentum, a powerful fundamental strategy, as a baseline for comparison.

Although recent frameworks such as FinCon (Yu YY et al., 2024) and FinAgent (Zhang WT et al., 2024) demonstrate strong portfolio optimization ability, they primarily focus on single-asset trading or portfolio construction. Their formulations do not naturally extend to multi-asset industry ranking; hence, we select SEP, StkFEP, and Momentum as more relevant baselines for our setting.

Additionally, to compare the performance of algorithms across different categories, we include the time-series model ARIMA (Shumway and Stoffer, 2019) and reinforcement learning algorithms—specifically, SAC, A2C, and PPO—from the FinRL framework (Liu XY et al., 2021) as baselines.

5.4 Main results

5.4.1 Performance on trending event detection (RQ1)

During the event detection phase, the results reveal that the LLM significantly surpasses clustering baselines across performance metrics (Table 1). Additionally, the O1 model outperforms the DeepSeek model in detecting trending events. Notably, the proposed event detection agent (EDA), combined with the optimal model and a memory-enhanced long-text reasoning workflow, achieves the highest accuracy.

Table 1 Comparison of models during event detection

Method	Accuracy	
	Financial news	KOL articles
Baseline	0.403	0.432
DeepSeek	0.535	0.581
O1	0.579	0.604
EDA	**0.650**	**0.683**

EDA: event detection agent. The best results are in bold

5.4.2 Performance on event prediction (RQ2)

For the event prediction task, as shown in Table 2, MENTOR outperforms the existing leading model in this field, Stk_FEP, on both the Chinese KOL articles and English financial news datasets, no matter whether Stk_FEP uses DeepSeek or O1 as its reasoning model. This superior performance is primarily due to MENTOR's ability to iteratively enhance its reasoning based on the reasoning system. Additionally, in this phase, the model simultaneously learns reasoning corrections for the next-stage ranking phase. By referencing the accuracy of the subsequent rankings, it can iteratively refine its reasoning. This allows it to predict a wider variety of financial events. We will further discuss this in Section 5.5.

Table 2 Comparison of model performance in the hot event prediction task on English financial news and Chinese KOL articles datasets

Method	Event prediction score*	
	Financial news	KOL articles
StkFEP_DeepSeek	4.06	3.90
StkFEP_O1	4.45	4.45
MENTOR_DeepSeek	4.50	**4.54**
MENTOR_O1	**4.55**	4.53

* Number of overlapping events between predicted and actual events. The best results are in bold

5.4.3 Performance on industry indices ranking (RQ3)

For the industry indices ranking task, as illustrated in Table 3, MENTOR outperforms the existing baseline models, SEP and the Momentum strategy, on both the S&P 500 and A-share datasets. It is noteworthy that even in S&P 500, where Momentum effects are highly pronounced, MENTOR's ranking performance surpasses that of the Momentum strategy.

On the A-share dataset, MENTOR O1 achieves a hit rate of 0.439, slightly smaller than that of DeepSeek's hit rate of 0.488. The statistical significance of the relevant results is detailed in the supplementary materials. Furthermore, MENTOR-DeepSeek outperforms O1, suggesting that DeepSeek may exhibit superior analytical capabilities when processing Chinese-language corpora.

5.5 Ablation study and incremental prediction value

Compared to SEP, the key differentiator of MENTOR is its introduction of multi-task decomposition for event prediction and ranking. Compared to StkFEP, MENTOR's main distinction lies in the introduction of the teacher–student interaction mechanism.

As demonstrated in Table 3, both modules added to MENTOR have a significant impact. Specifically, we find that: (1) The interaction mechanism can substantially improve the model's overall performance; (2) Multi-task decomposition, in turn, significantly enhances the optimization capability of

Table 3 Performance comparison of different methods on industry ranking tasks

Category	Method	S&P 500 (11 industries)			A-share (9 industries)		
		Spearman	Kendall_τ	Hit_rate_3	Spearman	Kendall_τ	Hit_rate_3
Rule-based	Momentum	0.159	0.120	0.364	0.100	0.065	0.444
Time-series	ARIMA	0.142	0.131	0.373	0.138	0.089	0.444
RL-based	SAC	0.089	0.067	0.303	0.092	0.071	0.389
	A2C	0.082	0.061	0.288	0.088	0.068	0.378
	PPO	0.095	0.073	0.318	0.097	0.075	0.401
LLM-based	StkFEP_DeepSeek	0.050	0.049	0.333	0.103	0.060	0.444
	StkFEP_O1	0.072	0.068	0.303	0.101	0.102	0.389
	SEP_DeepSeek	0.078	0.042	0.208	0.118	0.069	0.472
	SEP_O1	0.121	0.098	0.333	0.134	0.098	0.408
	MENTOR_DeepSeek (ours)	0.220	0.170	0.364	0.141	**0.128**	**0.488**
	MENTOR_O1 (ours)	**0.221**	**0.180**	**0.394**	**0.173**	0.125	0.439

Best results are in bold. RL: reinforcement learning; SAC: soft actor–critic; A2C: advantage actor–critic; PPO: proximal policy optimization

the interaction mechanism.

These findings collectively prove that our proposed MENTOR framework possesses broad applicability and strong generalization capabilities. Notably, we innovatively analyze and deconstruct the incremental value derived from making predictions. Traditional methods predominantly rely on deterministic reasoning based on known content, with predictions made only in the final step, as seen with FinCon. Despite complex processing of the inputs in the last step, assessing the reliability of the final decision-making is challenging. Our experiments achieve the same core objective as traditional ablation experiments through incremental prediction value analysis. We start with the outcomes of our predictions, breaking the prediction down into multiple verifiable sub-predictions, with each undergoing further analysis. Sub-prediction 1 involves the correlation between industry ranking outcomes and Momentum, and sub-prediction 2 examines the incremental value embodied in predicting event trends.

5.5.1 Incremental value in industry ranking

LLM-based forecasting models are prone to Momentum bias: in the absence of strong signals, they tend to extrapolate recent performance, and the narrative data on which they rely are by themselves skewed toward recently outperforming assets, which attract more coverage and bullish sentiment. In our experiments, a simple Momentum strategy consistently outperforms several LLM baselines, and the rankings produced by these models closely resemble Momentum-driven orderings. This suggests that much of their apparent predictive power may stem from exposure to Momentum rather than genuine forward-looking analysis. Thus, further investigation into the ranking portion and its incremental value relative to Momentum strategies is warranted.

As illustrated in Table 4, we have conducted a comparative analysis of similarity graphs between various algorithms and the Momentum algorithm. The experiments corroborate that the MENTOR framework exhibits the least correlation with Momentum, whereas other models display greater alignment with the Momentum effect. This validates that the MENTOR framework not only provides incremental value in ranking but also demonstrates a larger divergence from Momentum compared to other models, establishing it as the most Momentum-independent approach.

Table 4 Similarity between various LLM-based methods and the Momentum strategy on S&P 500

Method	Spearman	Kendall_τ	Hit_rate_3
StkFEP_DS	0.666	0.600	**0.556**
SEP_DS	0.602	0.576	0.593
MENTOR_DS	**0.563**	**0.527**	0.630

DS: DeepSeek. A lower similarity indicates a greater divergence between the predictions of the two methods, suggesting that textual information provides more diverse insights

5.5.2 Incremental value in predicting event trends

Previous experimental results have demonstrated the significant incremental value of the event prediction module. Consequently, we discuss herein the sources of this predictive increment. Our observations reveal that in the initial cycles, there is a high degree of repetition between the predicted news and the hotspots summarized from the original news; however, as the cycles progress, this value rapidly decreases to around 5–6, indicating that the iterative algorithm tends to drive the model towards making

diverse predictions.

Through our case study, we have observed that event prediction models tend to make divergent forecasts particularly when industry trends are poised for a reversal. For instance, despite the adverse effects of trade disputes on the computer sector, which experiences the largest decline within the week, the event prediction model anticipates heightened attention towards the computer industry in the subsequent week, rather than a continuation of the downward trend. Consequently, the ranking model adjusts its priorities and places the computer sector at the top of the list. In reality, the industry records the highest gains in the following week. Building upon our prior analysis of the ranking module's similarity to Momentum strategies, we can infer that such types of reversals are a primary source of deviation from Momentum algorithms. Evaluations of the overall forecast accuracy indicate that the possibility of successfully predicting these reversals is relatively high. This underscores the incremental value of the event prediction module, which is capable of offering perspectives that diverge from market consensus and is, in fact, a significant contributor to the generation of excess returns.

5.6 Scalability via Top K industry selection

To assess the scalability and robustness of our framework, we introduce a Top K industry selection mechanism. Let $R_t = \{i_1, i_2, \ldots, i_{|I|}\}$ denote the industry ranking generated by MENTOR at time t, where industries are ordered by their predicted narrative impact. Instead of evaluating the full ranking, we consider the truncated subset

$$\mathrm{TopK}_t = \{i_1, i_2, \ldots, i_K\}, \tag{11}$$

and restrict portfolio construction and downstream evaluation to this set. This allows us to test whether MENTOR's predictions remain effective under different levels of concentration and diversification, i.e., $K = 1$ (most concentrated), 3, and 5 (more diversified).

We analyze three key questions:

1. Can MENTOR sustain stable excess returns over a long horizon?

2. Where do the model's excess returns primarily come from?

3. Do different teacher–student choices affect the level and stability of excess returns?

Fig. 4 reports the cumulative excess returns of the Top K portfolios for the DeepSeek-R1 teacher and DeepSeek-V3 student combination.

We observe that the Top 3 portfolio delivers the highest excess returns, significantly outperforming both Top 5 and the market benchmark. This performance remains stable across the evaluation window, answering our first question on long-term consistency.

Regarding the second question, the superior performance of the most highly ranked industries suggests that excess returns are concentrated in the top signals, potentially benefiting from Momentum effects and the clearer narrative signals associated with leading industries. In contrast, industries ranked in the middle are less distinct and add limited incremental value when included in broader Top K

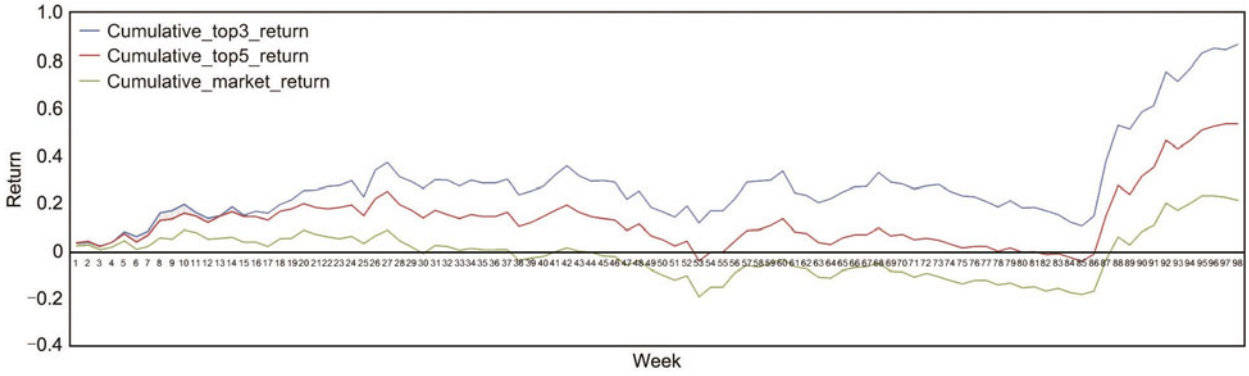

Fig. 4 Cumulative return of the Top K portfolios. The figure illustrates the backtesting results for the DeepSeek-R1 teacher and DeepSeek-V3 student pair; more results are reported in Section 5.6 (References to color refer to the online version of this figure)

Table 5 Performance of teacher–student configurations evaluated by Spearman correlation and excess return

Teacher model	Qwen3-30B-Instruct*		Qwen3-235B-Instruct*		DeepSeek-V3*	
	Spearman	Excess return (%)	Spearman	Excess return (%)	Spearman	Excess return (%)
DeepSeek-R1	0.132	27.51	0.176	40.48	0.141	45.80
Doubao-Seed	0.153	47.45	0.141	30.12	0.161	36.75
O1-Preview	0.151	38.21	0.122	29.69	0.173	42.74
O3	0.139	37.39	0.158	34.52	0.160	44.40
O4-Mini	0.127	21.38	0.152	46.98	0.164	47.99
Qwen3-30B-Think	0.120	17.53	0.144	33.78	0.146	31.61
Qwen3-235B-Think	0.147	36.98	0.168	34.48	0.166	43.51

* Student model

Table 6 Performance of the DeepSeek model across multiple investment metrics

Metric	Value
avg_top3_return	0.7063
avg_top5_return	0.4915
avg_market_return	0.2483
avg_excess_top3	0.4580
avg_excess_top5	0.2431
volatility_top3	3.7631
volatility_top5	3.3603
sharpe_top3	0.1217
sharpe_top5	0.0723
win_rate_top3	0.6122
win_rate_top5	0.6327
hit_ratio_top3	0.5068
hit_ratio_top5	0.4878
Total number of weeks	98

The reported indicators comprehensively evaluate return (average return, excess return), risk (volatility), and risk-adjusted performance (Sharpe ratio, hit ratio, and win rate)

portfolios.

For the third question, Table 5 summarizes the results across different teacher–student configurations. All combinations generate positive and statistically meaningful excess returns, though the magnitude and volatility vary. In particular, DeepSeek-V3 consistently yields the most stable and substantial gains, which is why we adopt it as the main benchmark in our overall evaluation.

In our evaluation, we adopt a comprehensive set of investment metrics to capture different dimensions of model performance in Table 6. Specifically, average returns and excess returns are employed to assess profitability relative to the market benchmark, while volatility measures the associated risk exposure. Sharpe ratios further provide a risk-adjusted evaluation of efficiency. Finally, win rates and hit ratios reflect the predictive accuracy of the model in identifying outperforming stocks. Together, these indicators offer a balanced view of both return and risk, enabling a robust assessment of investment effectiveness.

5.7 Efficiency and practical significance

Regarding the efficiency of the MENTOR framework, by employing a teacher–student collaborative architecture with large and small models, we reduce the token consumption of the inference model along with the associated computational and time costs. We achieve significant performance improvements with minimal increase in the number of inference calls. This approach not only effectively enhances efficiency in practical applications, but also provides a valuable reference for future research on collaborative work between think and instruct models.

6 Conclusions

We propose MENTOR, which has superior predictive performance across datasets encompassing multiple languages and diverse sources. Compared to existing models, including robust reasoning systems like O1, MENTOR not only surpasses them in performance metrics but also introduces methodological innovations in deconstructing and analyzing the incremental value of predictions. Empirical evaluations underscore MENTOR's enhanced capabilities. When applied to datasets from markets encompassing multiple industries, our model demonstrates significant improvements in forecasting accuracy over existing models. These results affirm that integrating advanced reasoning into predictive models increases the precision of event forecasting and yields more reliable predictions of industry index rankings. In conclusion, the MENTOR methodology effectively facilitates the transfer of sophisticated reasoning across models, amplifying predictive performance through reinforced reasoning functions. This advancement lays a robust foundation for future applications in financial forecasting and narrative

trend analysis, indicating that enhancing reasoning capabilities is pivotal for improving the reliability and accuracy of predictive models.

Contributors

Liyuan CHEN, Gaoguo JIA, and Dongsheng GU designed the research. Liyuan CHEN, Gaoguo JIA, Dongsheng GU, and Yuhang JIANG processed the data. Liyuan CHEN, Gaoguo JIA, Yuhang JIANG, and Jiangpeng YAN drafted the paper. Xiu LI and Xiaojun ZENG helped organize, revised, and finalized the paper.

Conflict of interest

Liyuan CHEN, Xiu LI, and Xiaojun ZENG are guest editors of the Special Feature on Theories and Applications of Financial Large Models of *Frontiers of Information Technology & Electronic Engineering*; they were not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Ariyo AA, Adewumi AO, Ayo CK, 2014. Stock price prediction using the ARIMA model. Proc 16th Int Conf on Computer Modelling and Simulation, p.106-112. https://doi.org/10.1109/UKSim.2014.67

Bao WZD, Cao YT, Yang Y, et al., 2025. Data-driven stock forecasting models based on neural networks: a review. *Inform Fus*, 113:102616. https://doi.org/10.1016/j.inffus.2024.102616

Chen JL, Xiao ST, Zhang PT, et al., 2024. BGE M3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. https://arxiv.org/abs/2402.03216

Eimann REA, 2008. Network Event Detection with Entropy Measures. PhD Dissertation, The University of Auckland, Auckland, New Zealand.

Enevoldsen KC, Chung I, Kerboua I, et al., 2025. MMTEB: massive multilingual text embedding benchmark. Proc 13th Int Conf on Learning Representations, p.1-57.

Gao YF, Xiong Y, Gao XY, et al., 2024. Retrieval-augmented generation for large language models: a survey. https://arxiv.org/abs/2312.10997

Guan Y, Peng H, Wang XZ, et al., 2024. OpenEP: open-ended future event prediction. https://arxiv.org/abs/2408.06578

Koa KJL, Ma YS, Ng R, et al., 2024. Learning to generate explainable stock predictions using self-reflective large language models. Proc ACM Web Conf, p.4304-4315. http://doi.org/10.1145/3589334.3645611

Li ML, Li S, Wang ZHL, et al., 2021. The future is not one-dimensional: complex event schema induction by graph modeling for event prediction. Proc Conf on Empirical Methods in Natural Language Processing, p.5203-5215. https://doi.org/10.18653/v1/2021.emnlp-main.422

Li QZ, Chao Y, Li D, et al., 2023. Event detection from social media stream: methods, datasets and opportunities. Proc IEEE Int Conf on Big Data, p.3509-3516. https://doi.org/10.1109/BigData55660.2022.10020411

Liang QQ, Zhu MY, Zheng XL, et al., 2021. An adaptive news-driven method for CVaR-sensitive online portfolio selection in non-stationary financial markets. Proc 13th Int Joint Conf on Artificial Intelligence, p.2708-2715. https://doi.org/10.24963/ijcai.2021/373

Liu MP, Zhu ML, Wang XY, et al., 2024. ECHO-GL: earnings calls-driven heterogeneous graph learning for stock movement prediction. Proc 38th AAAI Conf on Artificial Intelligence, p.13972-13980. https://doi.org/10.1609/aaai.v38i12.29305

Liu NF, Lin K, Hewitt J, et al., 2024. Lost in the middle: how language models use long contexts. *Trans Assoc Comput Ling*, 12:157-173. https://doi.org/10.1162/tacl_a_00638

Liu XY, Yang HY, Gao JC, et al., 2021. FinRL: deep reinforcement learning framework to automate trading in quantitative finance. Proc 2nd ACM Int Conf on AI in Finance, Article 1. https://doi.org/10.1145/3490354.3494366

Llama Team, 2024. The LLaMA 3 herd of models. https://arxiv.org/abs/2407.21783

Ma QC, Luo XF, Luo Y, 2014. Information entropy based the stability measure of user behaviour network in microblog. Proc 10th Int Conf on Semantics, Knowledge and Grids, p.67-74. https://doi.org/10.1109/SKG.2014.29

Mroua M, Lamine A, 2023. Financial time series prediction under Covid-19 pandemic crisis with long short-term memory (LSTM) network. *Human Soc Sci Commun*, 10:530. https://doi.org/10.1057/S41599-023-02042-W

Packer C, Fang V, Patil SG, et al., 2024. MemGPT: towards LLMs as operating systems. https://arxiv.org/abs/2310.08560

Qian H, Zhou HT, Zhao Q, et al., 2024. MDGNN: multi-relational dynamic graph neural network for comprehensive and dynamic stock investment prediction. Proc 38th AAAI Conf on Artificial Intelligence, p.14642-14650. https://doi.org/10.1609/aaai.v38i13.29381

Ratner N, Levine Y, Belinkov Y, et al., 2023. Parallel context windows for large language models. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.6383-6402. https://doi.org/10.18653/v1/2023.acl-long.352

Schubert E, Sander J, Ester M, et al., 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans Datab Syst*, 42(3):19. https://doi.org/10.1145/3068335

Shi Y, Wang YN, Qu Y, et al., 2024. Integrated GCN-LSTM stock prices movement prediction based on knowledge-incorporated graphs construction. *Int J Mach Learn Cybern*, 15(1):161-176. https://doi.org/10.1007/s13042-023-01817-6

Shiller RJ, 2019. Narrative Economics: How Stories Go Viral & Drive Major Economic Events. Princeton University Press, Princeton, USA.

Shumway RH, Stoffer DS, 2019. ARIMA models. In: Time Series (1st Ed.). Chapman and Hall/CRC.

Wang XD, Salmani M, Omidi P, et al., 2024. Beyond the limits: a survey of techniques to extend the context length in large language models. Proc 33rd Int Joint Conf on Artificial Intelligence, p.8299-8307. https://doi.org/10.24963/ijcai.2024/917

Xiao GX, Tian YD, Chen BD, et al., 2024. Efficient streaming language models with attention sinks. Proc 12th Int Conf on Learning Representations, p.1-13.

Yang HY, Liu XY, Wang CD, 2023. FinGPT: open-source financial large language models. https://arxiv.org/abs/2306.06031

Yang MY, Zheng XL, Liang QQ, et al., 2022. A smart trader for portfolio management based on normalizing flows. Proc 31st Int Joint Conf on Artificial Intelligence, p.4014-4021. https://doi.org/10.24963/ijcai.2022/557

Ye CC, Hu ZN, Deng YH, et al., 2024. MIRAI: evaluating LLM agents for event forecasting. https://arxiv.org/abs/2407.01231

Yu YY, Li HH, Chen Z, et al., 2023. FinMEM: a performance-enhanced LLM trading agent with layered memory and character design. Proc 3rd AAAI Spring Symp Series, p.595-597. https://doi.org/10.1609/aaaiss.v3i1.31290

Yu YY, Yao ZY, Li HH, et al., 2024. FinCon: a synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. Proc 38th Int Conf on Neural Information Processing Systems, Article 4354.

Yuksekgonul M, Bianchi F, Boen J, et al., 2024. TextGrad: automatic "differentiation" via text. https://arxiv.org/abs/2406.07496

Zhang WT, Zhao LX, Xia HC, et al., 2024. FinAgent: a multimodal foundation agent for financial trading: tool-augmented, diversified, and generalist. https://arxiv.org/abs/2402.18485v1

Zhang YZ, Li M, Long D, et al., 2025. Qwen3 embedding: advancing text embedding and reranking through foundation models. https://arxiv.org/abs/2506.05176

Zhao L, 2022. Event prediction in the big data era: a systematic survey. *ACM Comput Surv*, 54(5):94. https://doi.org/10.1145/3450287

Zheng LM, Chiang WL, Sheng Y, et al., 2023. Judging LLM-as-a-judge with MT-Bench and chatbot Arena. Proc 37th Int Conf on Neural Information Processing Systems, Article 2020.

Zhu FQ, Gao J, Yu CL, et al., 2023. A generative approach for script event prediction via contrastive fine-tuning. Proc 37th AAAI Conf on Artificial Intelligence, p.14056-14064. https://doi.org/10.1609/aaai.v37i11.26645

Zhu P, Li YT, Hu YF, et al., 2024. MCI-GRU: stock prediction model based on multi-head cross-attention and improved GRU. *Neurocomputing*, 638:130168. https://doi.org/10.1016/j.neucom.2025.130168

List of supplementary materials

1 Evaluation for trending event detection
2 Statistical significance of performance gains
3 Prompt templates
4 Data samples of Chinese KOL articles texts and English financial news texts