

FinRAQ: Financial News Reasoning through Importance Ranking and Socratic Questioning

Zee Hen Tang¹ Mi-Yen Yeh^{1,2}

¹Data Science Degree Program, National Taiwan University and Academia Sinica

²Institute of Information Science, Academia Sinica

Taipei, Taiwan

d11948008@ntu.edu.tw, miyen@iis.sinica.edu.tw

Abstract

In financial markets, effective trading decisions require accurately interpreting unstructured information, especially news articles. However, prior work overlooks the complex news and agent interactions when making trading decisions. To address these limitations, we propose FinRAQ, which enhances the news importance scoring by modeling semantic and temporal interactions between news articles. FinRAQ incorporates a Socratic questioning stage into the debate framework, prompting agents to challenge each other’s reasoning through Socratic questions. Extensive experiments across 21 assets show that FinRAQ outperforms state-of-the-art baselines, achieving up to 38% improvement in annualized returns.

1 Introduction

Financial markets react rapidly to information released through news articles, which reflect corporate developments, macroeconomic signals, and investor sentiment. Accurately interpreting such unstructured information is therefore critical for effective trading decisions. However, financial news reasoning remains challenging because relevant signals are often distributed across multiple articles, whose implications interact in complex, sometimes conflicting ways. A trading system must not only read individual news items but also reason about their relative importance and interactions.

Large language models (LLMs) have shown strong capabilities as financial decision-making tools (Li et al., 2023b; Wang et al., 2024c). Prior studies have shown that incorporating LLM-based news analysis into trading agents can improve trading performance (Zhang et al., 2024a; Kirtac and Germano, 2024), highlighting the potential of LLMs for financial reasoning.

Despite these advances, current methods exhibit several limitations. First, *lack of explicit news prioritization*. Many approaches concatenate multiple

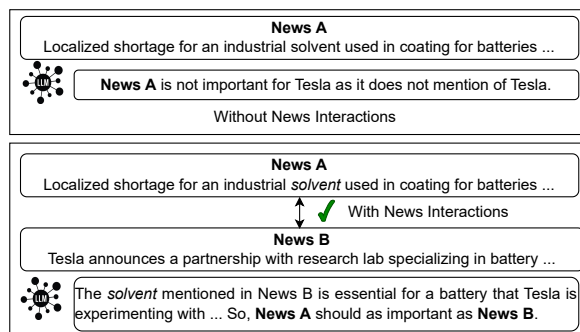


Figure 1: Lack of News Interactions.

news articles into a single long prompt, implicitly assuming that LLMs can identify the most relevant information without explicit guidance. In practice, the importance of news varies significantly across context, making this assumption unreliable. Second, *lack of modeling interactions* among news events. The significance of a news article often depends on its relationship with other events. As illustrated in Figure 1, a supply-chain disruption that appears unrelated to a company may become highly relevant when considered together with corporate partnerships or strategic announcements. Existing methods fail to explicitly capture such semantic and temporal interactions among news events, leading to inconsistent or misleading interpretations. Third, *limited agent interaction during reasoning*. To incorporate diverse viewpoints, some studies simulate bullish and bearish agents engaging in debate (Xing, 2025; Li et al., 2023b). However, these debates typically rely on simple argument exchanges, which often produce persuasive but shallow reasoning or unresolved conflicts. More importantly, existing debate frameworks rarely enforce structured reasoning among agents, making it difficult to expose flawed assumptions or inconsistencies in their arguments systematically.

To address these limitations, we propose FinRAQ, a financial reasoning framework that im-

proves trading decisions through structured news analysis and Socratic questioning. FinRAQ consists of two primary modules: a news importance scoring module and a Socratic debate module. For news prioritization, FinRAQ introduces NewsRank, a temporal graph-based model built upon Temporal Graph Networks (TGN) (Rossi et al., 2020). NewsRank incorporates a Dual Channel Message Passing (DC-MP) mechanism to capture both semantic and temporal interactions among news articles, enabling the system to estimate the importance of each news item dynamically. FinRAQ trains separate NewsRank models from bullish and bearish perspectives, allowing agents to reason about market sentiment from opposing viewpoints.

For agent reasoning, FinRAQ integrates Socratic questioning (SQ) into a multi-agent debate framework to enforce deeper reasoning between agents. Unlike prior work that primarily applies Socratic prompting to improve reasoning within a single LLM, FinRAQ leverages Socratic questioning to structure interactions among competing financial agents. During the debate process, agents challenge each other’s assumptions through structured Socratic questions that probe definitions, hypotheses, and logical consequences. This mechanism systematically exposes weak arguments and encourages agents to refine their reasoning, leading to more robust trading decisions.

Our contributions are summarized as follows: (i) **A news interaction-aware financial reasoning framework.** We propose FinRAQ, which explicitly models interactions among news events and integrates structured reasoning for LLM-based trading decisions. (ii) **Temporal graph-based news importance modeling.** We introduce NewsRank, a temporal graph model with Dual Channel Message Passing that captures both semantic and temporal interactions among news articles to estimate their importance dynamically. (iii) **Socratic questioning for multi-agent financial reasoning.** We incorporate structured Socratic questioning into a multi-agent debate framework to expose flawed reasoning and encourage deeper financial reasoning. (iv) **Extensive empirical evaluation on real-world markets.** Experiments on 21 assets across multiple sectors with four backbone LLMs show that FinRAQ consistently outperforms strong baselines across several financial metrics, achieving up to 38% improvement in annualized returns.

2 Related Work

2.1 LLM Agents for Decision-Making

Financial LLMs such as PIXIU (Xie et al., 2023), BloombergGPT (Wu et al., 2023), FinGPT (Yang et al., 2023), FinLlama (William Todt, 2023), and Instruct-FinGPT (Zhang et al., 2023) have shown strong trading performance through finetuning on domain-specific data. FinMem (Li et al., 2024) and FinAgent (Zhang et al., 2024b) further enhance trading performance through memory and reflection mechanisms. FinCon (Yu et al., 2024) and TradingAgents (Xiao et al., 2025) introduce multi-agent frameworks with risk control and role specialization. SEP (Koa et al., 2024) applies PPO with a self-reflective agent for autonomous stock reasoning. MarketSenseAI (Fatouros et al., 2024) focuses on generating interpretable investment insights from large-scale textual data. TiMi (Song et al., 2026) proposes a rationality-driven agentic system with a comprehensive multi-modal dataset for quantitative trading. News-based stock prediction methods such as PEN (Li et al., 2023a), FININ (Wang et al., 2024b), and LLMFactor (Wang et al., 2024a) model financial news for market prediction. However, these approaches focus on stock movement prediction without explicitly modeling news importance from contrasting market perspectives. Different from prior work, FinRAQ explicitly models news importance from bullish and bearish perspectives by capturing temporal and semantic news interactions for trading decisions.

2.2 Socratic Questioning in LLMs

Socratic questioning has been widely adopted across domains to enhance the reasoning capabilities of LLMs. Qi et al. (2023) propose a divide-and-conquer algorithm that mimics recursive human thinking to improve complex reasoning. Liu et al. (2024), Ding et al. (2024), and Bonino et al. (2024) leverage Socratic-style instructions to improve the teaching abilities of LLMs. For code debugging, Ashok Kumar and Lan (2024) and Kargupta et al. (2024) refine Socratic questions to improve model performance. More recently, Peng et al. (2025) incorporate a Socratic module into a multi-agent debate framework to combine Socratic questioning with divergent thinking. Distinct from prior work, FinRAQ introduces Socratic questioning into a financial multi-agent debate framework, generating more strategic questions that expose flawed reasoning for trading decisions.

3 FinRAQ Framework

FinRAQ employs two NewsRank models (§3.2), conditioned on bullish or bearish outlook, to assign dynamic importance scores to news articles, which guide the LLM agents in predicting stock movements. To reach consensus from conflicting predictions, FinRAQ integrates the Socratic questioning into debate (§3.3) to enhance the depth of reasoning underlying the trading decision. The overall framework is shown in Figure 2.

3.1 NewsRank

NewsRank incorporates a novel Dual Channel Message Passing (DC-MP) mechanism into a temporal graph network to refine the importance scores for each news item from bullish and bearish perspectives on a 1-5 scale, where the score reflects only the impact of a news article on stock movement, without indicating the direction of the movement. These dynamically updated scores allow NewsRank to adapt to evolving news contexts and effectively prioritize the most impactful news.

Initial Importance Score. The importance scores are designed to capture the potential impact of news on stock movement. FinRAQ assesses news influence from bullish and bearish perspectives by assigning two initial importance scores for each news item. For each score, FinRAQ prompts the LLM to evaluate the news under two hypothetical scenarios: one where the stock price is expected to rise, producing the bullish importance score, and another where the stock price is expected to fall, producing the bearish importance score. For each news article, NewsRank queries the LLM five times under each perspective using a 5-point Likert scale. The bullish and bearish importance scores are obtained by averaging the five responses, yielding a continuous score in the range [1, 5].

Temporal and Semantic Interactions. NewsRank models news flow as a temporal sequence of interaction events. For a news item n_i published at time t_i , NewsRank defines the interaction events with two types of neighbors: temporal neighbors P and semantic neighbors Q . $P = \{p_j | 0 < \Delta(t) < \tau\}$ comprises news items published within a temporal window of size τ , where $\Delta(t) = t_i - t_j$, $|P| = \mu$, and τ is the temporal window size. $Q = \{q_j | \text{sim}(n_i, q_j) > \kappa\}$ consists of news items with cosine similarity to n_i exceeds a similarity threshold κ , with $|Q| = \nu$. Each inter-

action event is represented as a tuple

$$\delta(t, n_i, n_j) = (v_i, v_j, e_{ij}, t) \quad \forall n_j \in P \cup Q, \quad (1)$$

where v_i and v_j are the nodes for n_i and n_j , respectively, and e_{ij} is the edge feature computed as the cosine similarity between their news embeddings.

For each interaction event, NewsRank adopts the same update function as TimeSGN (Xu et al., 2024) to update the state vector $s_i(t)$ and $s_j(t)$ for node v_i and v_j . Moreover, FinRAQ maintains an importance score memory which records the refined score $\hat{y}_i$, computed by Eq. (12), to capture the dynamic importance of each news. Details of the state update function are provided in Appendix B.

Dual Channel Message Passing. NewsRank introduces DC-MP to generate the temporal embeddings $z_i^\ell(t)$ of node v_i at time t in layer ℓ , capturing temporal and semantic information in news interactions. For simplicity, we omit the timestep notation (t) and use z_i^ℓ to represent $z_i^\ell(t)$.

The input to the ℓ -th layer consists of node v_i 's embedding $z_i^{\ell-1}$ in the previous layer, the embeddings of v_i 's temporal and semantic neighbors, the timestamps, edge features, and refined importance scores. Suppose $Z_P^\ell = [z_{p_1}^{\ell-1} \parallel \dots \parallel z_{p_\mu}^{\ell-1}]$ and $Z_Q^\ell = [z_{q_1}^{\ell-1} \parallel \dots \parallel z_{q_\nu}^{\ell-1}]$ denote the temporal embeddings of P and Q in $(\ell - 1)$ -th layer. $T_P = \{t_{p_1}, \dots, t_{p_\mu}\}$ and $T_Q = \{t_{q_1}, \dots, t_{q_\nu}\}$ denote the timestamps of P and Q . $E_P = [e_{i,p_1} \parallel \dots \parallel e_{i,p_\mu}]$ and $E_Q = [e_{i,q_1} \parallel \dots \parallel e_{i,q_\nu}]$ are the edge features. $Y_P = [\hat{y}_{p_1} \parallel \dots \parallel \hat{y}_{p_\mu}]$ and $Y_Q = [\hat{y}_{q_1} \parallel \dots \parallel \hat{y}_{q_\nu}]$ are the refined scores of P and Q . $\parallel$ is the concatenation operator and the temporal embeddings z_i^ℓ are calculated as

$$K_P^t = V_P^t = [\phi(t - t_{p_1}) \parallel \dots \parallel \phi(t - t_{p_\mu})], \quad (2)$$

$$K_Q^t = V_Q^t = [\phi(t - t_{q_1}) \parallel \dots \parallel \phi(t - t_{q_\nu})], \quad (3)$$

$$q_P^t = \text{MA}_P^t(q^t, K_P^t, V_P^t), \quad (4)$$

$$q_Q^t = \text{MA}_Q^t(q^t, K_Q^t, V_Q^t), \quad (5)$$

$$q_P^\ell = \text{MA}_P^\ell(q^\ell, [Z_P^{\ell-1} \parallel E_P \parallel Y_P], [Z_P^{\ell-1} \parallel E_P \parallel Y_P]), \quad (6)$$

$$q_Q^\ell = \text{MA}_Q^\ell(q^\ell, [Z_Q^{\ell-1} \parallel E_Q \parallel Y_Q], [Z_Q^{\ell-1} \parallel E_Q \parallel Y_Q]), \quad (7)$$

$$h_P = \sigma(\text{MLP}_P(q_P^\ell \parallel q_P^t)), \quad (8)$$

$$h_Q = \sigma(\text{MLP}_Q(q_Q^\ell \parallel q_Q^t)), \quad (9)$$

$$\alpha = \psi(\text{MLP}_\alpha(h_P \parallel h_Q)), \quad (10)$$

$$z_i^\ell(t) = \alpha h_P + (1 - \alpha) h_Q + z_i^{\ell-1}(t), \quad (11)$$

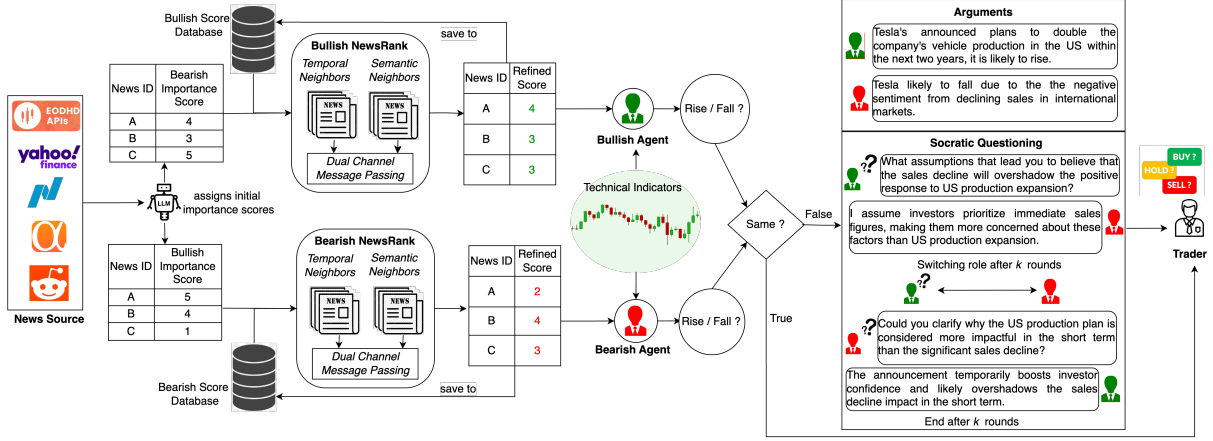


Figure 2: Overview Framework of FinRAQ, consisting of the NewsRank module for importance scoring and the Socratic questioning module for debate-based trading decisions.

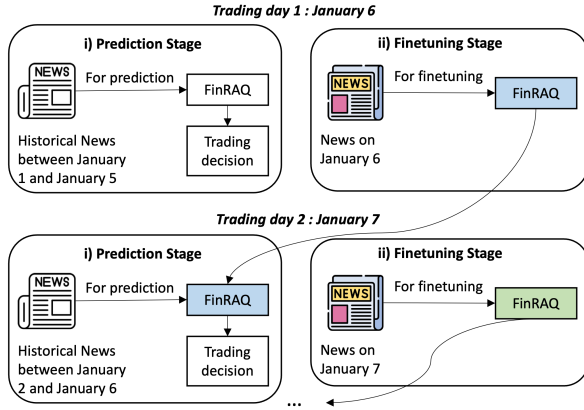


Figure 3: Daily trading procedure of FinRAQ.

where MA_P^l , MA_Q^l , MA_P^t , MA_Q^t denote the multi-head attention layers, ϕ is the time-encoding function (da Xu et al., 2020), σ refers to the GELU function, and ψ is the sigmoid function. The queries are defined as $q^l = z_i^{l-1}(t)$, $q^t = \phi(0)$. $z_i^0(t) = s_i(t) + f_i(t)$, where $f_i(t)$ is the node feature of v_i . DC-MP first aggregates the news information from temporal neighbors P and semantic neighbors Q separately. The temporal context h_P and semantic context h_Q are fused via a gated mechanism α into the updated temporal embeddings $z_i^l(t)$, enabling NewsRank to dynamically combine temporal and semantic contexts. Given the temporal embeddings of the last layer $z_i^{L-1}(t)$, the refined importance score $\hat{y}_i$ is computed as

$$\hat{y}_i = (1 + 4 \cdot \psi(W z_i^{L-1}(t) + b)), \quad (12)$$

where W and b are trainable parameters and the refined score is normalized to the range [1-5]. We store each news item with its corresponding score

in a database.

Objective Functions. Instead of relying on absolute importance scores, NewsRank adopts a ranking loss $\mathcal{L}_R$ to model the relative importance of news articles within the same temporal window, as relative rankings are more informative than absolute scores for news prioritization. A consistency loss $\mathcal{L}_C$ further ensures that semantically similar and temporally proximate news items have consistent temporal embeddings, preventing NewsRank from assigning inconsistent importance scores to similar news items that could lead to contradictory LLM interpretations. $\mathcal{L}_R$ and $\mathcal{L}_C$ are defined as

$$\mathcal{L}_R = \frac{1}{|P|} \sum_{n_j \in P} \max(0, -\mathbb{I}_{y_i > y_j} (\hat{y}_i - \hat{y}_j) + 1), \quad (13)$$

$$\mathcal{L}_C = \frac{\sum_{n_j \in P \cap Q} \max(0, \gamma - \text{sim}(z_i^l(t), z_j^l(t)))}{|P \cap Q|}, \quad (14)$$

where y_i and y_j are the initial scores for news items n_i and n_j , respectively, and κ denotes the similarity threshold. The final loss function $\mathcal{L}$ is defined as $\mathcal{L} = \mathcal{L}_R + \mathcal{L}_C$.

3.2 Training and Inference

For each target asset, FinRAQ trains two NewsRank models from bullish and bearish perspectives using news from both the target asset and its peer assets. The NewsRank is trained once and finetuned daily to incorporate the latest news. On average, the initial training requires around 27 minutes for 6,868 news articles, while daily finetuning requires around 7 minutes for 47 news articles on an NVIDIA A6000 GPU, making FinRAQ highly

practical for real-world daily trading. Figure 3 illustrates the two-stage daily trading procedure of FinRAQ: (i) a prediction stage, where news articles from the preceding five days are used to make a trading decision, and (ii) a finetuning stage, where NewsRank is updated using the current day’s news to adapt to the latest market conditions.

3.3 Socratic Questioning

The bullish and bearish agent analyze news articles using refined importance scores assigned by their respective NewsRank models and technical indicators to predict stock movements. When predictions conflict, FinRAQ initiates a debate using SQ to elicit deeper reasoning. SQ compels each agent to address the opponent’s Socratic questions that expose flaws or vague reasoning, leading to more robust reasoning. Following prior work (Chang, 2023), FinRAQ defines six types of questions for SQ: Definition, Hypothetical, Elenchus, Maieutic, Reductio, and Dialectic. Each category is designed to probe the underlying logic of an argument and promote deeper analytical reasoning. In each round, one agent acts as questioner, selecting a question type and posing a Socratic question based on the other’s arguments and responses. The respondent is required to answer the question directly. After k rounds, the roles are switched, and the new questioner repeats the questioning process. Once both agents complete their questioning turns, an LLM trader makes the trading decision based on each agent’s reasoning, including how effectively each agent identifies weaknesses in the opponent’s arguments, the clarity of their responses, and whether any logical inconsistencies exist.

We compare the trading decisions made by FinRAQ with and without SQ during the debate process. A trading decision is considered correct if it aligns with the actual stock price movement on the following day (i.e., buying before a price increase or selling before a price decrease). If a trader makes an incorrect decision without SQ but makes a correct one when SQ is applied, it suggests that SQ effectively improves trading decisions. To learn from these improvements, FinRAQ finetunes an LLaMA-3.3-70B model using question-answer pairs from these successful cases in the training period, enabling the model to learn question patterns that expose weak reasoning and generate more strategic questions. During inference, FinRAQ applies the finetuned model to generate more strategic Socratic questions, enhancing the overall quality of the de-

bate process. Details of the Socratic question types and finetuning are provided in Appendix B.

4 Experimental Setup

We conduct experiments on 21 assets across different sectors and market capitalizations. Following Xiao et al. (2025), we construct a multi-modal dataset for each asset from diverse sources. Each dataset includes news articles, financial reports, and technical indicators. We compare FinRAQ with two rule-based baselines (B&H and MACD), three news-driven methods (PEN, FININ, LLMFactor), and five LLM-based baselines (FinMem, FinAgent, FinCon, TradingAgents (TA), and TiMi). We evaluate FinRAQ across four different backbone LLMs, including GPT-5.4, Claude-Sonnet-4.6, Gemini-3-Pro, and Llama-4-Maverick-17B-128E-Instruct. We evaluate all methods using three metrics: (i) Annualized Return (AR) for overall performance, (ii) Sharpe Ratio (SR) for risk-adjusted returns, and (iii) Maximum Drawdown (MDD) for worst-case loss. FinRAQ is trained on news from March 1, 2022, to March 1, 2024, and evaluated from March 1, 2024, to March 1, 2026. Trading decisions are made daily using up to 50 news items from the past five days ($\tau = 5$), with short selling allowed. All experiments are repeated eight times, and we report the average results. Full experimental details are provided in Appendix A.

5 Experimental Results

Our experiments aim to answer the following research question. **RQ1:** Does FinRAQ consistently outperform state-of-the-art baselines across sectors and market capitalizations? **RQ2:** Does Socratic Questioning outperform other prompting strategies in generating more effective agent interactions for trading decisions? **RQ3:** What is the individual contribution of each component in FinRAQ to its overall trading performance? **RQ4:** What are the distributions of importance score changes and Socratic question types in FinRAQ? **RQ5:** How do the number of peer assets and SQ rounds affect FinRAQ’s performance? **RQ6:** Is FinRAQ practical for daily trading deployment?

For all results below, we conduct the Wilcoxon signed-rank test with Holm-Bonferroni correction to assess statistical significance.

	NVDA			JPM			JNJ		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	47.2	1.0	36.9	30.3	1.1	24.4	27.8	1.4	14.5
MACD	7.2	0.3	41.3	13.6	0.7	18.7	22.2	1.4	9.3
PEN	-15.2	-0.1	59.0	4.3	0.2	27.1	3.3	0.2	19.4
FININ	-11.6	0.0	63.4	7.6	0.3	26.7	7.6	0.4	27.0
LLMFactor	6.7	0.3	48.4	37.0	1.3	19.8	14.5	0.7	15.9
FinMem	40.7	0.9	42.7	52.4	1.7	16.2	24.1	1.2	11.8
FinAgent	45.0	1.0	42.7	54.7	1.8	18.3	40.2	1.9	21.2
FinCon	71.4	1.3	43.5	71.8	2.2	16.0	52.4	2.4	10.1
TA	85.5	1.5	48.3	92.4	2.7	22.3	63.8	2.8	15.9
TiMi	91.3	1.5	39.4	80.4	2.4	16.3	61.7	2.7	12.6
FinRAQ	129.5	1.9	31.8	115.0	3.1	9.4	89.5	3.6	9.1
	RIG*			LRN*			BTC-USD		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	12.9	0.5	68.1	17.7	0.6	64.1	2.5	0.2	49.7
MACD	26.1	0.7	39.4	18.1	0.7	57.7	0.6	0.1	32.5
PEN	-3.8	0.2	48.5	9.0	0.5	70.0	-17.4	-0.2	58.8
FININ	1.6	0.3	61.0	2.9	0.4	58.5	-3.3	0.1	44.0
LLMFactor	14.3	0.5	64.6	36.1	0.8	40.5	48.0	1.0	40.0
FinMem	42.4	0.9	55.3	40.7	0.9	74.3	56.8	1.1	44.3
FinAgent	50.3	1.0	40.2	55.6	1.0	33.2	79.2	1.4	37.2
FinCon	54.1	1.0	52.7	52.5	1.0	45.1	81.1	1.4	36.3
TA	68.7	1.2	34.4	76.9	1.3	59.0	94.1	1.6	28.2
TiMi	77.6	1.3	59.3	65.7	1.1	43.8	102.2	1.7	21.8
FinRAQ	97.2	1.4	28.7	87.4	1.4	30.7	128.4	1.9	16.0

Table 1: Performance comparison across sectors using GPT-5.4 ($p < 0.05$). '*' denotes mid-cap stocks (2–10 billion USD) selected from the Russell 2000 index. Best results are highlighted in bold.

	NVDA			JPM			JNJ		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
Period	2025-11-10 ~2026-02-06			2025-02-13 ~2025-05-12			2025-03-31 ~2025-06-26		
B&H	-25.4	-0.7	14.1	-20.2	-0.5	24.4	-27.7	-1.2	11.8
TiMi	58.7	1.5	9.6	106.6	2.1	11.3	67.8	1.9	7.0
FinRAQ	76.6	1.7	9.2	154.6	2.2	9.4	84.0	2.5	6.6
	RIG*			LRN*			BTC-USD		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
Period	2025-04-02 ~2025-06-30			2025-10-28 ~2026-01-26			2025-12-30 ~2026-02-28		
B&H	85.9	1.4	23.1	96.9	1.6	12.4	-68.3	-2.0	35.3
TiMi	91.7	1.4	23.2	158.3	2.2	7.0	80.6	1.4	13.4
FinRAQ	108.9	1.6	21.6	209.6	3.0	6.4	93.5	1.5	11.8

Table 2: Performance comparison under different market conditions ($p < 0.05$). Red and green represent bearish and bullish periods, respectively.

5.1 Performance Comparison (RQ1 & RQ2)

Table 1 shows that FinRAQ consistently outperforms all baselines across all assets. LLM-based baselines generally outperform traditional rule-based and ML&RL methods, demonstrating the

advantage of LLMs in processing complex financial information. Although LLMFactor improves the explainability of trading decisions, it still underperforms other LLM-based baselines, likely due to the lack of explicit modeling of news importance.

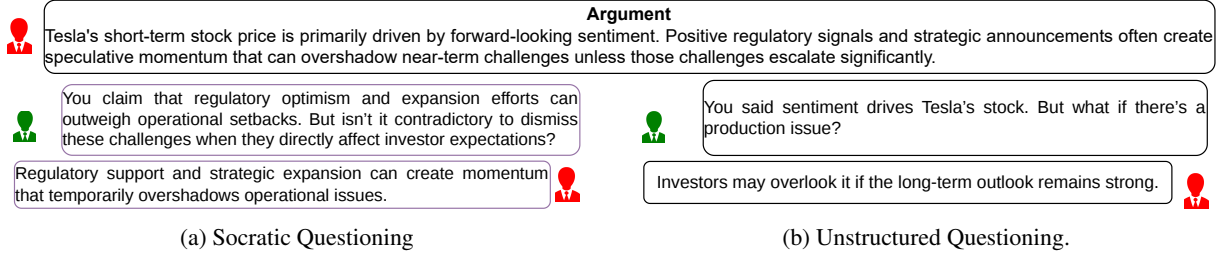


Figure 4: Comparison between Socratic questioning and unstructured questioning.

	AR%↑	SR↑	MDD%↓
CoT	94.6	1.6	32.9
Self-consistency	95.9	1.6	32.4
Reflexion	96.6	1.6	46.9
MAD	98.5	1.6	53.5
UQ	100.1	1.6	45.2
FinRAQ	114.6	1.8	32.0

Table 3: Performance comparison between different prompting strategies on NVDA ($p < 0.05$).

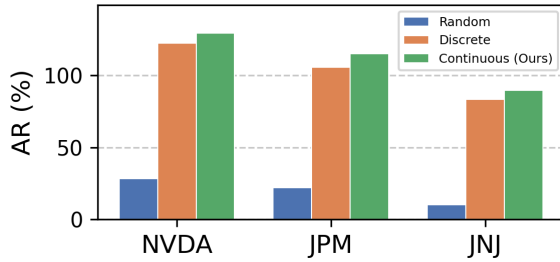
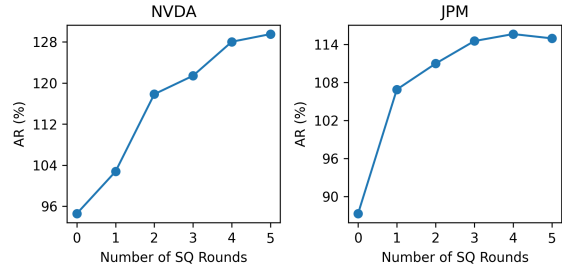
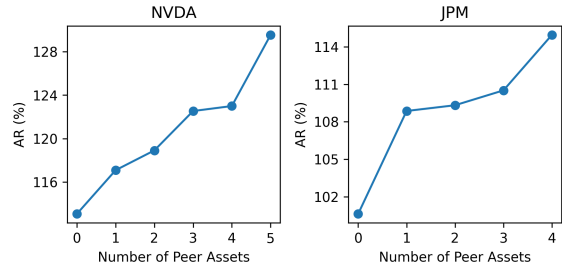


Figure 5: Comparison of different scoring methods.



(a) Effects of SQ Rounds on AR.



(b) Effects of Peer Assets on AR.

Figure 6: Effect of SQ rounds and peer assets on AR.

FinRAQ explicitly models the importance of news and challenges weak reasoning during the decision-making process, achieving the best performance across all metrics and improving performance by up to 38% on NVDA. FinRAQ also outperforms all baselines on mid-cap stocks, such as RIG and LAUR, indicating strong robustness across different market capitalizations.

To evaluate different market conditions, we select testing periods with B&H AR above 20% and below 20% to represent generally bullish and bearish markets, respectively. Table 2 shows that FinRAQ outperforms the baselines under both bullish and bearish market conditions, demonstrating robust performance across diverse market environments. FinRAQ achieves the lowest standard deviation across all metrics and consistently outperforms all baselines across four backbone LLMs. We also conduct experiments using a testing period beyond the knowledge cutoff of GPT-4o to

investigate whether the performance gains of FinRAQ are driven by information from the model’s prior knowledge. Detailed results across different backbone LLMs and market conditions, as well as evaluations beyond the knowledge cutoff, are provided in Appendix C.1.

To evaluate the effectiveness of SQ, we compare FinRAQ with five prompting strategies, including Chain-of-Thought (CoT) (Wei et al., 2022), Self-Consistency (Wang et al., 2023), Reflexion (Shinn et al., 2023), Multi-agent Debate (MAD) (Liang et al., 2024), and Unstructured Questioning (UQ). Unlike SQ, UQ allows agents to pose questions without structural constraints, serving as a control to assess the value of structured Socratic questioning. Since all baselines are training-free, the finetuned question generator is removed from FinRAQ to ensure a fair comparison. Table 3 shows that FinRAQ consistently outperforms all prompting strategies, demonstrating the effectiveness of

	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	45.8	1.0	44.2
w/o $\mathcal{L}_R$	58.0	1.1	40.2
w/o $\mathcal{L}_C$	112.8	1.7	35.5
w/o Finetuning	114.6	1.8	32.0
FinRAQ	129.5	1.9	31.8

Table 4: Component-wise ablation results on NVDA.

SQ in identifying weaknesses in arguments. Appendix C.2 provides detailed results for all assets. Figure 4 compares the results of SQ and UQ. While UQ merely raises a broad concern, such as "what if there's a production issue?", SQ examines the internal logic of the arguments and explicitly confronts potential contradictions. SQ encourages agents to challenge conflicting reasoning, revealing limitations in sentiment-driven valuations and reducing the influence of persuasive but flawed arguments.

5.2 Ablation Studies (RQ3)

To evaluate the contribution of each component in FinRAQ, we conduct an ablation study by removing: (i) the NewsRank module (w/o NR), (ii) the ranking loss (w/o $\mathcal{L}_R$) (iii) the consistency loss (w/o $\mathcal{L}_C$), and (iv) the fine-tuned LLaMA model for generating Socratic questions (w/o FT). As shown in Table 4, removing any component degrades overall performance, underscoring the importance of each component. Removing NewsRank leads to the most significant degradation, emphasizing the importance of dynamically prioritizing news through semantic and temporal modeling. Removing $\mathcal{L}_R$ leads to a larger performance drop than removing $\mathcal{L}_C$, indicating that the ranking task contributes more to improving trading performance than the consistency task. The fine-tuned LLaMA model further improves the quality of Socratic questions, yielding up to a 21% improvement in AR for NVDA. To evaluate the effectiveness of our continuous importance scoring, we compare it against two alternatives: (i) random assignment and (ii) discrete scoring via majority voting. Figure 5 shows that the continuous importance score achieves the best performance, likely because it captures more nuanced differences in news importance, leading to more accurate news prioritization. Comprehensive ablation studies, including robustness analysis of initial importance scores across different LLMs and the gating mechanism of NewsRank are provided in Appendices C.3–C.5.

5.3 NewsRank and SQ Analysis (RQ4)

Importance Score Changes: We analyze the importance score changes induced by both Bullish and Bearish NewsRank to quantify NewsRank’s influence on news importance. We define the average importance score change over 21 assets and N trading days as:

$$\frac{1}{\sum_{i=1}^{21} N_i} \sum_{i=1}^{21} \sum_{j=1}^{N_i} |y_{i,j} - \hat{y}_{i,j}|$$

where N_i is the total number of news articles for the i -th asset during the testing period, while $y_{i,j}$ and $\hat{y}_{i,j}$ denote the original and refined importance scores of the j -th news article of the i -th asset, respectively. As shown in Table 5, approximately half of all news articles have a score change in the range of (1, 2], indicating moderate refinement in most cases. Both bullish and bearish NewsRank exhibit similar overall trends, with only about 10% of articles showing a score change exceeding 2, and no article exceeding 4. However, Bearish NewsRank shows slightly more extreme cases, with a higher proportion of score changes exceeding 3 and a higher proportion of unchanged scores (i.e., score change equal to 0). These results suggest that NewsRank refines importance scores conservatively, avoiding abrupt changes that could introduce noise into trading signals. We also provide a case study of bearish NewsRank refines news importance scores through semantic interactions, demonstrating that NewsRank effectively leverages past news articles to refine the importance of recent news articles. Detailed analysis is provided in Appendix C.7.

Socratic Question Types: Due to computational constraints, we randomly sample 150 conversations with correct trading decisions and 150 conversations with incorrect trading decisions for each asset, resulting in 6,300 conversations in total (300 per asset). Given the substantial agreement between Gemini-2.5-Pro and human annotations ($\kappa = 0.86$), we apply Gemini-2.5-Pro to label all 6,300 conversations. We then evaluate the proportion of each question type in correct and incorrect conversations to identify which question types are more likely to lead to correct trading decisions (see Appendix C.6 for details of the κ computation). Table 6 shows that Elenchus achieves the highest correct rate, suggesting that directly challenging the core assumptions effectively uncovers flawed reasoning and enhance the trading decisions. In contrast, Dialectic

	0	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	Score Change
Bullish	17.4% $\pm$ 0.6	25.9% $\pm$ 2.2	46.0% $\pm$ 4.3	6.8% $\pm$ 0.6	3.9% $\pm$ 0.5	0% $\pm$ 0.0	1.6 $\pm$ 0.3
Bearish	21.6% $\pm$ 1.2	15.2% $\pm$ 3.6	51.4% $\pm$ 4.7	7.1% $\pm$ 0.4	4.7% $\pm$ 0.3	0% $\pm$ 0.0	1.9 $\pm$ 0.5

Table 5: Percentage distribution of importance score changes and mean score change after refinement, averaged over 21 assets across all trading days in the testing period.

	C (%)	IC (%)	Frequency
Definition	44	56	747
Hypothetical	55	45	1,470
Elenchus	63	37	2,064
Maieutic	52	48	906
Reductio	67	33	651
Dialectic	37	68	462

Table 6: Correct (C) and incorrect (IC) trading decision rates across different question types.

	Latency (mins)	Input Tokens
FinAgent	5.4 $\pm$ 0.6	12,746 $\pm$ 543
FinCon	9.6 $\pm$ 1.2	14,468 $\pm$ 301
TA	13.7 $\pm$ 2.3	16,482 $\pm$ 585
TiMi	9.4 $\pm$ 1.6	9,661 $\pm$ 292
FinRAQ	12.3 $\pm$ 0.5	9,481 $\pm$ 258

Table 7: Average per-inference costs over 21 assets during the testing period.

achieves the lowest correct rate, possibly because synthesizing opposing viewpoints introduces ambiguity that confuses the trader rather than clarifying the decision. Both Definition and Dialectic yield correct rates below 50% and exhibit lower occurrence, suggesting that LLMs may struggle to apply these two question types to enhance reasoning.

5.4 SQ Rounds and Peer Assets (RQ5)

Figure 6a shows that increasing the number of SQ rounds leads to higher AR, indicating that deeper agent interactions improve the decision quality. The improvement is marginal beyond 4 rounds, suggesting diminishing returns. This is likely because excessive questioning can introduce confusion, redundancy, or misaligned reasoning among agents. Figure 6b shows that incorporating news from more peer assets further improves AR, demonstrating the importance of capturing news interactions across related assets. We provide detailed results in Figure 15 and 16.

5.5 Inference Efficiency (RQ6)

We evaluate the per-inference efficiency costs of FinRAQ to assess its practical deployability for daily trading. Table 7 reports that FinRAQ consumes fewer input tokens than most baselines, as the majority of its LLM calls are used for Socratic Questioning, which takes only the previous conversation history as input, without long news articles or technical indicators. While FinRAQ’s inference latency includes both the latency of LLM calls and daily NewsRank finetuning, the total cost of approximately 12 minutes per day remains practical for daily trading, where only one trading decision is made. Compared with the substantial performance gains shown in Table 1, the inference latency is an acceptable trade-off, demonstrating that FinRAQ achieves deployment efficiency for practical daily trading. Detailed results are provided in Table 45.

6 Conclusion

We propose FinRAQ, a financial trading framework that introduces NewsRank to capture broader market signals from bullish and bearish perspectives. Furthermore, FinRAQ integrates Socratic questioning into the multi-agent debate framework, with question type selection further improved through finetuning. Extensive experiments on 21 assets across five sectors demonstrate that FinRAQ consistently outperforms state-of-the-art baselines across diverse market conditions. Future work will explore memory mechanisms to enhance the Socratic questioning process further.

7 Limitations

One limitation of our work is the computational cost of training NewsRank, which may limit its applicability in high-frequency trading. However, since FinRAQ is designed for daily trading, the initial training is a one-time process. Once trained, NewsRank is finetuned daily to incorporate the latest news, taking approximately 10 minutes, making it practical for real-world daily trading applications. Although four LLMs demonstrate high agreement in assigning importance scores ($\alpha = 0.82$),

it would be valuable to examine whether similar agreement holds for smaller language models. Evaluating the robustness of FinRAQ across models of different scales remains a promising direction for future work.

8 Ethical Considerations

All data used in this study were collected from publicly available sources, including financial news articles, historical stock prices, and technical indicators. The data are used solely for academic research purposes. We ensure that our use of data complies with the terms of service of each data provider and does not infringe upon any intellectual property or privacy rights. No proprietary or private information is utilized.

Additionally, the models developed in our work are not intended for real-world deployment without proper regulatory review and validation. We explicitly discourage any misuse of the proposed system for unauthorized financial trading or speculative activities.

9 Acknowledgement

This work was supported in part by National Science and Technology Council, Taiwan, R.O.C., under Grant 113-2628-E-001-003-MY4; the National Taiwan University and Academia Sinica Innovative Joint Program, under grant AS-NTU-114-06; the Institute of Information Science, Research Center for IT Innovation, Academia Sinica.

References

- Nischal Ashok Kumar and Andrew Lan. 2024. [Improving socratic question generation using data augmentation and preference optimization](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 108–118, Mexico City, Mexico. Association for Computational Linguistics.
- Giulia Bonino, Gabriele Sanmartino, Giovanni Gatti Pinheiro, Paolo Papotti, Raphael Troncy, and Pietro Michiardi. 2024. Euler: Fine tuning a large language model for socratic interactions. In *Proceedings of the Second International Workshop on Artificial Intelligence Systems in Education co-located with 23rd International Conference of the Italian Association for Artificial Intelligence (AIXIA 2024)*. CEUR Workshop Proceedings.
- Edward Y. Chang. 2023. [Prompting large language models with the socratic method](#). Preprint, arXiv:2303.08769.
- da Xu, chuanwei ruan, evren korpeoglu, sushant kumar, and kannan achan. 2020. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*.
- Yuyang Ding, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He. 2024. [Boosting large language models with socratic method for conversational mathematics teaching](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 3730–3735, New York, NY, USA. Association for Computing Machinery.
- George Fatouros, Kostas Metaxas, John Soldatos, and Dimosthenis Kyriazis. 2024. [Can large language models beat wall street? evaluating gpt-4's impact on financial decision-making with marketsenseai](#). *Neural Computing and Applications*.
- Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani Tur, and Jiawei Han. 2024. [Instruct, not assist: LLM-based multi-turn planning and hierarchical questioning for socratic code debugging](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9475–9495, Miami, Florida, USA. Association for Computational Linguistics.
- Kemal Kirtac and Guido Germano. 2024. [Sentiment trading with large language models](#). *Finance Research Letters*, 62:105227.
- Kelvin J.L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. [Learning to generate explainable stock predictions using self-reflective large language models](#). In *The Web Conference 2024*.
- Haohang Li, Yangyang Yu, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Su-chow, and Khaldoun Khashanah. 2024. [Finmem: A performance-enhanced LLM trading agent with layered memory and character design](#). In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- Shuqi Li, Weiheng Liao, Yuhan Chen, and Rui Yan. 2023a. [Pen: prediction-explanation network to forecast stock price movement with better explainability](#). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23*. AAAI Press.
- Yang Li, Yangyang Yu, Haohang Li, Zhi Chen, and Khaldoun Khashanah. 2023b. [Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance](#). Preprint, arXiv:2309.03736.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages

- 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.
- Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. [SocraticLM: Exploring socratic personalized teaching with large language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Xian Peng, Pan Yuan, Dong Li, Junlong Cheng, Qin Fang, and Zhi Liu. 2025. [KELE: A multi-agent framework for structured socratic teaching with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16342–16362, Suzhou, China. Association for Computational Linguistics.
- Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. [The art of SOCRATIC QUESTIONING: Recursive thinking with large language models](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. In *ICML 2020 Workshop on Graph Representation Learning*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. [Reflexion: language agents with verbal reinforcement learning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Zifan Song, Kaitao Song, Guosheng Hu, Ding Qi, Junyao Gao, Xiaohua Wang, Dongsheng Li, and Cairong Zhao. 2026. [Trade in minutes! rationality-driven agentic system for quantitative financial trading](#). In *The Fourteenth International Conference on Learning Representations*.
- Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024a. [LLMFactor: Extracting profitable factors through prompts for explainable stock movement prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3120–3131, Bangkok, Thailand. Association for Computational Linguistics.
- Mengyu Wang, Shay B Cohen, and Tiejun Ma. 2024b. [Modeling news interactions and influence for financial market prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3302–3314, Miami, Florida, USA. Association for Computational Linguistics.
- Saizhuo Wang, Hang Yuan, Lionel M. Ni, and Jian Guo. 2024c. [Quantagent: Seeking holy grail in trading by self-improving large language model](#). *Preprint*, arXiv:2402.03755.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Pedram Babaei William Todt, Ramtin Babaei. 2023. Fin-llama: Efficient finetuning of quantized llms for finance. <https://github.com/Bavest/fin-llama>.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kam-badur, David Rosenberg, and Gideon Mann. 2023. [Bloomberggpt: A large language model for finance](#). *Preprint*, arXiv:2303.17564.
- Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2025. [Tradingagents: Multi-agents LLM financial trading framework](#). In *The First MARW: Multi-Agent AI in the Real World Workshop at AAAI 2025*.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. [Pixiu: A large language model, instruction data and evaluation benchmark for finance](#). *Preprint*, arXiv:2306.05443.
- Frank Xing. 2025. [Designing heterogeneous llm agents for financial sentiment analysis](#). *ACM Trans. Manage. Inf. Syst.*, 16(1).
- Yuanyuan Xu, Wenjie Zhang, Ying Zhang, Maria Or-lowska, and Xuemin Lin. 2024. [TimeSGN: Scalable and Effective Temporal Graph Neural Network](#). In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 3297–3310, Los Alamitos, CA, USA. IEEE Computer Society.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. [Fingpt: Open-source financial large language models](#). *FinLLM Symposium at IJCAI 2023*.
- Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, GUO-JUN XIONG, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. [Fincon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. [Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models](#). *Preprint*, arXiv:2306.12659.

Haohan Zhang, Fengrui Hua, Chengjin Xu, Hao Kong, Ruiting Zuo, and Jian Guo. 2024a. [Unveiling the potential of sentiment: Can large language models predict chinese stock price movements?](#) *Preprint*, arXiv:2306.14222.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. 2024b. [A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist](#). In *KDD*, pages 4314–4325.

A Experimental Setup

A.1 Datasets

We conduct experiments on 21 assets from five different sectors. Table 8 lists each asset and its corresponding sector. For each sector, we select two assets from the S&P 500 with market capitalizations exceeding \$200 billion and two assets from the Russell 2000 with market capitalizations between \$2 billion and \$10 billion, ensuring broad market coverage for evaluation.

For each asset, FinRAQ also incorporates news from its peer assets to enrich the news context and capture the sophisticated news interactions. To identify peer assets for each ticker, we query three LLMs (GPT-5.4, Claude-Sonnet-4.6, and Gemini-3-Pro), each prompted to recommend the top-10 peer assets based on industry classification and market capitalization. We retain only the assets unanimously recommended by all three LLMs as the final candidate peer assets. If more than five unanimous assets are identified, we rank them by their average ranking across the three LLMs and retain only the top-5. A human expert then reviews the candidate peer assets to verify their relevance, and any irrelevant assets are removed. Table 9 lists the peer assets for each ticker, which are selected based on their industry classification and market capitalization to ensure relevance and diversity in the news context. Following Xiao et al. (2025), we construct a comprehensive multi-modal dataset for each asset (including its peer assets), consisting of five components:

(i) *News Articles*: Daily news collected from diverse sources, including Yahoo Finance, EODHD, Nasdaq, Reddit, and Seeking Alpha, covering both company-specific developments and broader macroeconomic trends. The macroeconomic news includes coverage related to politics, Federal Reserve rate decisions, U.S. Treasury bonds, major U.S. stock market indices, European and Japanese central banks, and foreign exchange rates.

(ii) *Financial Reports*: Companies’ quarterly and annual filings, such as Form 10-K and Form 10-Q. A total of 16 reports are collected for each asset across the training and testing periods.

(iii) *Historical Stock Prices*: Daily open, high, low, close, and adjusted close prices for each asset.

(iv) *Company Profiles and Financial History*: Descriptions of company profiles, industry focus, and historical financial information provided by third-party sources.

(v) *Technical Indicators*: Sixty standard technical analysis indicators computed for each asset, including RSI, MACD, SMA, and KDJ.

Table 10 shows the dataset statistics.

A.2 Baseline Models

We evaluate FinRAQ against 2 rule-based, 3 news-driven models, and 5 LLM-based models:

Buy and Hold (B&H): B&H is a long-term investment strategy where an asset is purchased and held regardless of market fluctuations. It assumes markets will rise over time, aiming to capture long-term capital returns.

Moving Average Convergence Divergence (MACD): MACD tracks the difference between short-term and long-term moving averages to identify trend direction and momentum. Buy or sell signals are generated when the MACD line crosses above or below the signal line.

PEN (Li et al., 2023a): proposes a neural framework that jointly predicts stock price movements and generates explanations by identifying key driving factors, improving both prediction accuracy and interpretability in financial forecasting.

FININ (Wang et al., 2024b): FININ models the interactions among news items and their influence on market prices for stock movement prediction.

LLMFactor (Wang et al., 2024a): LLMFactor is a financial trading model that leverages large language models to analyze market data and generate trading signals. It integrates natural language processing capabilities to interpret news and financial reports, thereby enhancing trading decisions.

FinMEM (Li et al., 2024): FinMEM is a financial trading model that incorporates a memory module to store and retrieve relevant information from past market data and news. It uses the memory to inform trading decisions, allowing it to adapt to changing market conditions.

FinAgent (Zhang et al., 2024b) (Zhang et al., 2024b): FinAgent is a multimodal foundational

Sector	Assets
Technology	NVDA, AAPL, CIFR*, QBTS*
Financial	JPM, MS, MARA*, WULF*
Healthcare	JNJ, UNH, IOVA*, IBRX*
Energy	XOM, CVX, RIG*, PTEN*
Consumer Staples	WMT, KO, LAUR*, LRN*
Cryptocurrency	BTC-USD

Table 8: Assets and their corresponding sectors. '*' denotes mid-cap companies selected from the Russell 2000 index.

Asset	Peer Assets
NVDA	TSM, MSFT, AMD, AVGO, AAPL
AAPL	NVDA, MSFT, GOOGL, AMZN, META
JPM	GS, MS, BAC, C
MS	GS, JPM, BAC, BLK, C
JNJ	PFE, MRK, ABT, LLY
UNH	CVS, CI, HUM, CNC, ELV
XOM	CVX, COP, TTE
CVX	XOM, COP, TTE
WMT	COST, TGT, AMZN, DG, DLTR
KO	PEP, MNST, KDP, CCH
CIFR	IREN, WULF, RIOT, CORZ
QBTS	IONQ, RGTI, QUBT
MARA	RIOT, CORZ, IREN, CLSK, WULF
WULF	RIOT, MARA, CIFR, HUT, CLSK
IOVA	IBRX, MRK, BNTX
IBRX	IOVA, MRK, BNTX
RIG	VAL, NE, SDRL, DO, HAL
PTEN	HP, NBR, HAL, LBRT, RIG
LAUR	ATGE, STRA, LOPE, LRN, GHC
LRN	LOPE, STRA, ATGE, CHGG, LAUR
BTC-USD	MSTR, COIN, ETH-USD, MARA, BLK

Table 9: Peer assets for each target asset.

agent enhanced with tool augmentation for financial trading, capable of analyzing numerical, textual, and visual data through its market intelligence module.

FinCon (Yu et al., 2024): FinCon is an LLM-based multi-agent framework designed for financial trading. It incorporates dual-level risk control, allowing agents to refine investment strategies to improve trading decisions through self-critique.

TradingAgents (Xiao et al., 2025) (Xiao et al., 2025): TradingAgents is an LLM-based multi-agent framework that simulates a collaborative trading firm using LLM-powered agents in specialized roles, including analysts, risk managers, and traders with varying risk profiles.

TiMi (Song et al., 2026): TiMi proposes an agen-

tic trading framework where multiple LLM-driven agents follow rational decision-making principles to analyze financial information and coordinate trading decisions rapidly.

We apply GPT-5.4 as the backbone LLM for all LLM-based methods.

A.3 Evaluation Metrics

We compare FinRAQ with all baseline models using three metrics:

Annualized Return Rate (AR) measures the yearly profit or loss of an investment, indicating its overall performance. ARR is defined as:

$$ARR = \left(\left(\frac{V_T}{V_0} \right)^{\frac{N}{T}} - 1 \right) \times 100\%, \quad (15)$$

Asset	Total number of news	Average number of news per day
NVDA	16,304	5.58
AAPL	16,990	5.82
JPM	12,382	4.24
MS	15,394	5.27
JNJ	11,998	4.11
UNH	10,748	3.68
XOM	16,760	5.74
CVX	11,966	4.10
WMT	16,028	5.49
KO	10,670	3.65
CIFR	7,352	2.52
QBTS	5,218	1.79
MARA	15,314	5.24
WULF	6,452	2.21
IOVA	8,360	2.86
IBRX	7,116	2.44
RIG	7,766	2.66
PTEN	6,182	2.12
LAUR	4,636	1.59
LRN	4,376	1.50
BTC-USD	16,624	5.69

Table 10: Dataset Statistics for each asset.

where T is the total number of trading days, $N = 252$ is the number of trading days within a year. V_0 and V_T represent the initial and final portfolio values.

Sharpe Ratio (SR) measures the risk-adjusted return of an investment by comparing its excess return to its volatility. A higher Sharpe Ratio indicates a better risk-reward trade-off. SR is defined as:

$$SR = \frac{R_p - R_f}{\sigma_p}, \quad (16)$$

where R_p is the expected portfolio return, $R_f = 0.05$ denotes the risk-free rate, σ_p is the standard deviation of portfolio return.

Maximum Drawdown (MDD) captures the most significant decline from a peak to its lowest point, reflecting the worst-case scenario of investment decline, and helps assess the downside risk over a specific period.

A.4 Implementation Details

We train FinRAQ using news articles from March 1, 2022 to March 1, 2024, and evaluate the model from March 1, 2024 to March 1, 2026. We employ the NV-Embed-v2 model to generate news embeddings and compute the cosine similarity between pairs of news articles.

For the NewsRank module, we configure the number of layers $L = 2$, the similarity threshold κ is 0.7, and apply L2 regularizer. We train each NewsRank model for five epochs using the AdamW (Loshchilov and Hutter, 2019) optimizer, with a batch size of 32 and a learning rate of $5e-4$. The trading decisions are made daily based on news from the preceding five days ($\tau = 5$), with a maximum of 50 news items for each trading decision. Following prior work, short positions are allowed when the trader makes a *sell* decision. Following (Xiao et al., 2025), we set the annualized risk-free rate to 5%, based on the average yield of the 3-month U.S. Treasury bill during the evaluation period. Transaction costs, including slippage, short-selling borrowing fees, and commissions, are also set to be consistent with prior work. We implement FinRAQ in PyTorch and use publicly available implementations for baseline models whenever possible. For baselines without released code, we reimplement them based on the methodologies described in their respective papers.

B Methodology Details

B.1 State Update function

For each $\delta(t, n_i, g_j)$, NewsRank computes two messages to update the state vectors $s_i(t)$ and $s_j(t)$, of nodes v_i and v_j , respectively:

$$m_i = \text{msg}_i(s_i(t^-), s_j(t^-), x_i, x_j, \Delta t, e_{ij}), \quad (17)$$

$$m_j = \text{msg}_j(s_j(t^-), s_i(t^-), x_j, x_i, \Delta t, e_{ij}), \quad (18)$$

where msg_i and msg_j are the learnable message functions, x_i and x_j denote the news embeddings of nodes v_i and v_j , $s_i(t^-)$ and $s_j(t^-)$ are their previous state vectors. $s_i(t^-)$ encodes the temporal interactions of n_i with historical news items prior to time t and is updated after each interaction event. Given the previous state vectors $s_i(t^-)$ and $s_j(t^-)$, the messages m_i and m_j , the state update function is represented as

$$s_i(t) = s_i(t^-) + \text{LN}(\text{FFN}(m_i(t) \parallel \phi(t - t^-))), \quad (19)$$

$$s_j(t) = s_j(t^-) + \text{LN}(\text{FFN}(m_j(t) \parallel \phi(t - t^-))), \quad (20)$$

where ϕ is a generic time encoding function, $\text{LN}(\cdot)$ is the layer normalization, $\text{FFN}(\cdot)$ is a single feed-forward layer.

B.2 Socratic Questions Types

Following Chang (2023), FinRAQ applies six types of Socratic Questions:

- i. Definition: Clarify and explain the meaning of key terms and concepts.
 - Example: When you say AAPL is overvalued, which specific metric are you using to support that claim?
- ii. Hypothetical: Test the reasoning under different conditions, useful for understanding whether the argument is context-dependent or robust across various situations.
 - If AAPL were to announce a major product innovation next week, would that change your bearish outlook?
- iii. Elenchus: Uncover contradictions or inconsistencies and challenge the internal coherence of the reasoning.
 - Example: You rely on technical indicators for your forecast, yet you discount market sentiment. Isn't that contradictory?

iv. Maieutic: Drawing out the unstated assumptions or beliefs to prompt the respondent to examine their reasoning. Maieutic is inspired by the Socratic method of helping others "give birth" to their insights.

- Example: Why do you place so much weight on short-term earnings surprises? Isn't the market often pricing in future expectations?

v. Reductio: Extend an argument to its logical extreme to test its validity, effective for exposing overly simplistic views.

- Example: If past price action is always predictive, are you implying that markets are fully deterministic?

vi. Dialectic: Exploring opposing viewpoints to arrive at a more comprehensive understanding of a subject.

- Example: You emphasize valuation concerns, while I highlight momentum strength, might both perspectives reflect different market timeframes?

B.3 Finetuning LLaMA

Suppose the sequence of Socratic question-answer pairs is denoted as $\{q_1, a_1, \dots, q_k, a_k\}$, where k is the number of Socratic questioning (SQ) rounds, q_i represents the Socratic question posed in the i -th round, and a_i refers to the opponent's response in i -th round. FinRAQ fine-tunes the LLaMA model M using LoRA in a sequential-to-sequence formulation, optimizing the following objective:

$$L_{seq} = \sum_{i=1}^k \log p(q_i | q_1, a_1, \dots, q_{i-1}, a_{i-1}). \quad (21)$$

1,312 question-answer pairs are collected and used to fine-tune the model across five epochs. The LoRA configuration uses a rank of 8 and an alpha parameter of 32, with a batch size of 32 and a learning rate of $5e-4$.

C Full Results

C.1 Performance Comparison

Tables 11–16 present the performance of all models. FinRAQ consistently outperforms all baselines

across all metrics, including on mid-cap stocks, indicating strong robustness across sectors and market capitalizations. FinRAQ also achieves the lowest standard deviation across all metrics, demonstrating its stability in financial decision-making. FinRAQ consistently outperforms all news-driven baselines, likely due to two key limitations of existing approaches:

- (i) **Lack of LLM integration.** PEN and FININ do not incorporate LLMs for trading decisions. Prior work, such as FinCon (Yu et al., 2024), TradingAgents (Xiao et al., 2025), and TiMi (Song et al., 2026) has demonstrated that LLMs significantly improve financial trading performance when applied in a structured manner.
- (ii) **Lack of explicit news prioritization.** PEN and LLMFactor treat news articles or factors without explicit importance ranking. Without modeling news prioritization, it is difficult for these models to capture the relative impact of news on stock movements, particularly given the complexity and large volume of financial news.

These tables also compare FinRAQ’s performance with four backbone LLMs, including GPT-5.4 (GP), Claude-Sonnet-4.6 (CL), Gemini-3-Pro (GE), and Llama-4-Maverick-17B-128E-Instruct (LL). FinRAQ with different LLM backbones consistently outperforms all baselines, further demonstrating the robustness of our framework across LLM architectures. FinRAQ (LL) underperforms FinRAQ with GP, CL, and GE backbones across all metrics, likely due to its smaller model size and weaker reasoning capabilities compared to GPT-5.4, Claude-Sonnet-4.6, and Gemini-3-Pro. GPT-5.4, Claude-Sonnet-4.6, and Gemini-3-Pro achieve similar performance across all assets, indicating that they have comparable reasoning capabilities in financial decision-making. Due to API and computational constraints, only GPT-5.4 results are reported in the following experiments.

Robustness across different market conditions. Selecting a testing period that clearly represents a bullish or bearish market for all assets is challenging, as B&H AR varies across assets. For assets with B&H AR below 20% in Tables 11–16, we select an alternative testing period of at least 60 days where B&H AR exceeds 20% to approximate a bullish market environment. Conversely,

for assets with B&H AR above -20%, we select an alternative testing period of at least 60 days where B&H AR is below 20% to approximate a bearish market environment. Tables 17 and 18 show results under bullish market conditions, while Tables 19–24 show results under bearish market conditions. FinRAQ consistently outperforms all baselines in both scenarios, demonstrating its robustness across different market environments.

Evaluation beyond the Knowledge Cutoff. Table 25 presents the results using GPT-4o across four assets from different sectors. The testing period spans from July 1, 2024, to May 1, 2025, which is after GPT-4o’s knowledge cutoff date, thereby preventing information leakage from the testing period. FinRAQ outperforms all baselines, achieving up to a 41% improvement in annualized returns. These results suggest that the performance gains of FinRAQ are primarily attributable to the effectiveness of NewsRank and Socratic Questioning, rather than to the model’s prior knowledge of information from the testing period.

C.2 Prompting Strategies Comparison

To evaluate the effectiveness of SQ, we compare FinRAQ with the other 5 prompting strategies:

(i) **Chain-of-Thought (CoT)** (Wei et al., 2022): CoT prompting encourages the model to generate intermediate reasoning steps before providing a final answer, enhancing the model’s ability to solve complex problems.

(ii) **Self-Consistency (SC)** (Wang et al., 2023): SC prompting generates multiple reasoning paths and selects the most consistent answer among them, improving the reliability of the model’s responses.

(iii) **Reflexion** (Shinn et al., 2023): Reflexion prompting allows the model to reflect on its previous answers and reasoning steps, enabling it to identify and correct errors in its responses.

(iv) **Multi-agent Debate (MAD)** (Liang et al., 2024): MAD prompting involves multiple agents debating different perspectives on a given question, fostering critical thinking and diverse viewpoints to arrive at a more informed answer. The difference between MAD and FinRAQ is that MAD does not use SQ to guide the debate, while FinRAQ uses SQ to facilitate deeper interactions between agents.

(v) **Unstructured Questioning (UQ)**: UQ prompting involves asking a series of unstructured questions without following a specific format or strategy, allowing for more open-ended interactions but potentially lacking the depth and focus

provided by structured questioning methods like SQ.

Tables 26–31 show that FinRAQ with SQ consistently outperforms all other prompting strategies across all metrics, demonstrating the effectiveness of SQ in enhancing LLMs’ reasoning capabilities for financial decision-making. Figure 4 compares the results of SQ and UQ. While UQ merely raises a broad concern, such as "what if there’s a production issue?", SQ examines the internal logic of the arguments and explicitly confronts potential contradictions. SQ leads the respondent to acknowledge that sentiment-driven valuations can be overextended, revealing an inherent limitation. By requiring agents to challenge conflicting reasoning, SQ mitigates the risk of traders being influenced by persuasive but flawed reasoning.

C.3 Ablation Studies

Tables 32–37 demonstrate that removing the NewsRank module leads to a significant decline in AR, indicating the importance of accurately identifying key news for financial decision-making. The fine-tuned LLaMA further improves the quality of Socratic questions, resulting in more reliable trading decisions.

C.4 Analysis of Importance Score

To examine the design of the initial importance score, we compare two variants for assigning importance scores:

- (i) **Random:** Randomly assign an importance score between 1 and 5 to each news article.
- (ii) **Discrete:** Apply majority voting over the five LLM responses to determine the final importance score. Since the LLM assigns scores based on a 5-point Likert scale, majority voting produces a discrete score.

Different Scoring Methods. Tables 38–43 show that the continuous importance score outperforms both the random and discrete variants, demonstrating the effectiveness of our scoring design in capturing the nuanced importance of news articles. The random variant performs worst, likely because randomly assigned scores fail to reflect the relative importance of news. The discrete variant performs better than the random variant but still underperforms the continuous scores. This is likely because continuous scores enable more fine-grained differentiation among news articles, allowing the model to prioritize relevant information for trading decisions better.

Different Likert Scale. Figure 17 compares the granularity of various Likert scales (2, 3, 5, 7) for assigning importance scores. We observe that the improvement is marginal beyond 5 points, likely due to LLMs’ limited ability to consistently distinguish between adjacent score levels beyond 5, which introduces noise into the importance scores and diminishes the benefits of finer granularity.

Robustness across different LLMs. We prompt four LLMs (GPT-5.4, Claude-Sonnet-4.6, Gemini-3-Pro, and Llama-4-Maverick-17B-128E-Instruct) to independently assign importance scores to each news item, and compute Krippendorff’s α across the four scores to measure inter-LLM agreement. The resulting $\alpha = 0.82$ (95% CI: [0.78, 0.86]) indicates substantial agreement among the four LLMs in their importance score assignments, demonstrating the robustness of our scoring design across different LLM architectures.

C.5 Gating Mechanism Analysis

To study the effect of temporal and semantic interactions in NewsRank, we conduct an ablation study comparing different gating mechanisms in Equation (11). Each variant of α is defined as follows:

1. $\alpha = 0$: Semantic channel only, where the temporal channel is removed.
2. $\alpha = 1$: Temporal channel only, where the semantic channel is removed.
3. $\alpha = 0.5$: Average fusion of both semantic and temporal channels.
4. Learnable α : Dynamic gating as defined in Equations (11).

As shown in Tables 44, removing the temporal channel ($\alpha = 0$) leads to the largest performance drop, as FinRAQ loses the ability to capture recent market signals, since recent news often carries stronger predictive signals for short-term price movements than older historical events. Removing the semantic channel ($\alpha = 1$) also degrades performance, as FinRAQ can no longer retrieve historically similar news events that provide important context for interpreting current market reactions. Average fusion ($\alpha = 0.5$) improves over both single-channel variants, indicating that the two channels carry complementary information. However, it still underperforms the learnable variant, as treating both channels equally ignores the varying

relative importance of recency and historical similarity across different market conditions. FinRAQ dynamically balances the temporal and semantic channels via a learnable α based on the current market context, enabling more adaptive and effective news integration.

These results suggest that both temporal and semantic interactions contribute meaningfully to NewsRank, and that the learnable α is essential for effectively integrating them.

C.6 Analysis of Socratic Questions

We randomly sample 210 debate conversations across all assets (ten per asset). For each debate conversation, three annotators recruited from Prolific were tasked with manually identifying the questions that the trader emphasized most when making the final trading decision. Before starting the task, all annotators were provided with a consent form (Figure 13) for the collection and use of their annotations for research purposes. Figure 14 presents the instructions given to the annotators. Annotators were compensated above the local minimum wage to ensure fair payment.

We then apply Gemini-2.5-Pro to perform the same labeling task and measure its agreement with human annotations. The Cohen's Kappa score is $\kappa = 0.86$, indicating substantial agreement between humans and the LLM. Figure 12 shows the prompt for the annotation task.

C.7 Qualitative Analysis

Figure 18 presents a case study of bearish NewsRank refining news importance scores through semantic interactions. News A (2024/11/22) and News B (2024/11/25) report on the UK's CMA and India's CCI investigations into Apple, receiving refined importance scores of 4.6 and 2.9, respectively. On 2025/03/12, two new articles appear: News C, related to the CMA investigation, and News D, related to the CCI investigation. Both News A and News B are sampled as semantic neighbors by bearish NewsRank when refining the importance scores of News C and News D, respectively.

Although the original importance scores of News C and News D are 3.1 and 4.0, NewsRank refines them to 4.7 and 2.4, respectively. The refined scores reflect the influence of their historically similar counterparts, News A and News B, through semantic interactions, suggesting that NewsRank propagates the relative importance of historical news to semantically similar recent articles.

To validate this effect, we remove News A and News B from the dataset and recompute the scores. Without semantic interactions, the refined scores of News C and News D revert to 3.2 and 4.2, closely resembling their original unrefined scores. This contrast demonstrates that NewsRank effectively leverages past news articles to refine the importance of recent news articles.

D LLM Prompts

Figure 7 illustrates the prompt used to assign the importance scores to news articles, while Figure 8 presents the prompt for predicting stock movements. Figure 9 and 10 depict the prompts for generating Socratic questions and responses, respectively. Figure 11 provides the decision-making prompt for the LLM-based trading agent. We use AAPL as the illustrative stock example in all prompts.

E The Use of Large Language Models (LLMs)

LLMs were used solely for grammar correction and writing enhancement, and were not involved in the research design or methodology development.

Given the projected stock price trend for AAPL over the upcoming week (either bullish or bearish), your task is to carefully read and analyze each piece of news based on your expertise, investment strategy, style, and focus areas. Label the importance of each piece of news with a score <X> based on the following criteria:

1 - Irrelevant: The news has no impact on AAPL's stock price trend in the next week and can be ignored.

2 - Low Impact: The news has little relevance to the stock price, and ignoring it has minimal consequences.

3 - Moderate Impact: The news has some influence but is not a key driver of stock price changes.

4 - Highly Relevant: The news is likely to significantly impact the stock price, influencing its rise or fall.

5 - Extremely Important: The news has a decisive effect on the stock price trend, making it a critical factor for next week's movement.

(Note: Timing matters. Recent news is generally more impactful than older items.)

Start with a paragraph explaining how you determine the importance of each piece of news. Then, provide a concise summary of each news item, followed by a sentence indicating the importance of the news with a score labeled <X> (where X is a number between 1 and 5).

Example Response:

[Provide a paragraph explaining which types of news are important and which are not, along with the reasoning.]

Here are the AAPL-related news items from YYYY-MM-DD to YYYY-MM-DD (ordered from most recent to least recent):

YYYY-MM-DD

[One-sentence summary of the first news item]. (One sentence explaining the importance of the first news item) <X>

[One-sentence summary of the second news item]. (One sentence explaining the importance of the second news item) <X>

YYYY-MM-DD

...continue until the last AAPL-related news item.

Here are the global economic developments from YYYY-MM-DD to YYYY-MM-DD (ordered from most recent to least recent):

YYYY-MM-DD

[One-sentence summary of the first news item]. (One sentence explaining the importance of the first news item) <X>

[One-sentence summary of the second news item]. (One sentence explaining the importance of the second news item) <X>

YYYY-MM-DD

...continue until the last global economic news item.

Analyze each news article and explain its importance to AAPL, assigning an importance score to each news article. Please make sure the number of news in each day of the response is the same as the input. Do not fabricate any news or provide summaries and opinions beyond the specified structure. Please do not skip any news and make sure each news has an importance score.

Figure 7: Prompt for rating the importance of AAPL's news using a 1–5 importance scale.

You are a stock market expert specializing in analyzing news related to AAPL. Your task is to predict whether AAPL's stock price will rise or fall over the next few days based on:

1. The provided news articles.
2. The importance scores assigned to each news piece.

Each piece of news will be appended with an importance score <X> ranging from 1 to 5.

This score reflects the level of impact the news has on AAPL's stock price trend:

- 1 - Irrelevant: The news has no impact on AAPL's stock price trend in the next few days and can be ignored.
- 2 - Low Impact: The news has little relevance to the stock price, and ignoring it has minimal consequences.
- 3 - Moderate Impact: The news has some influence but is not a key driver of stock price changes.
- 4 - Highly Relevant: The news is likely to significantly impact the stock price, influencing its rise or fall.
- 5 - Extremely Important: The news has a decisive effect on the stock price trend, making it a critical factor for the next few days' movement.

Analyze the provided news and their importance scores to predict whether AAPL will rise or fall over the next few days. Your prediction should include a clear, logical explanation that references the assigned importance scores and highlights the key factors influencing your conclusion.

Response Format:

Your response should adhere to the following format:

Reasoning: (Provide a brief explanation of your reasoning, incorporating importance scores and key insights from your analysis.)

Predicted stock price movement for the next few days: (Only answer "Rise" or "Fall")

Additional Guidelines:

- Avoid fabricating information or deviating from the specified format.
- Ensure your analysis and prediction are based solely on the provided data.
- Your response should be clear, concise, and grounded in logical reasoning, demonstrating expertise in financial analysis and stock price prediction.

Figure 8: Prompt for predicting AAPL's stock movements.

You are a financial expert participating in a structured Socratic debate on the short-term price movement of AAPL. Your goal is to critically examine your opponent's reasoning through clear, focused responses and strategic questioning. You will be provided with:

1. Your Original Arguments: Your initial stance and prediction regarding AAPL's short-term stock price movement.
2. Opponent's Rebuttal: A critique of your original argument, highlighting weaknesses or flaws.
3. Opponent's Arguments: Your opponent's prediction on AAPL's short-term stock price movement, predicting the opposite of your stance.
4. Questioning History: A transcript of all questions and answers exchanged during the debate rounds.

Debate Format:

The debate consists of 4 rounds between you and your opponent.

Each round includes two actions:

1. Thought – Clearly formulate a questioning strategy aimed at uncovering flaws, inconsistencies, or hidden assumptions in your opponent's argument or responses. Maintain a coherent line of inquiry throughout the rounds. For each round, build upon their answers with deeper, logically connected follow-up questions that probe further into their logic, evidence, or assumptions.
2. Ask – Based on your strategy, pose a question using one of the six Socratic question types (Definition, Hypothetical, Elenchus, Maieutic, Reductio, or Dialectic). Your question should challenge your opponent's position and be informed by their previous response. Ensure your questions progressively develop your line of inquiry, creating a thoughtful and consistent progression across rounds.

There are six Socratic questions types (select one per round):

1. Definition: Clarify ambiguous terms or metrics in their argument.

Example: "When you say AAPL is 'overvalued,' what metric are you using to support that?"

2. Hypothetical: Test their reasoning under alternative scenarios.

Example: "If AAPL announces a major product innovation next week, would that alter your bearish stance?"

3. Elenchus: Expose internal contradictions in their reasoning.

Example: "You rely on technical indicators for your prediction, yet dismiss market sentiment. Isn't that inconsistent?"

4. Maieutic: Uncover unspoken assumptions or beliefs.

Example: "Why do you trust short-term earnings surprises to dictate stock trends? Isn't the market often forward-looking?"

5. Reductio: Push their logic to an extreme or absurd outcome.

Example: "If you claim past price action is always predictive, are you suggesting markets are entirely deterministic?"

6. Dialectic: Synthesize opposing views to find deeper truth.

Example: "You cite valuation headwinds, I emphasize momentum strength—could both explain different market timeframes?"

Output Format:

Your Thought: [Clearly express your reasoning or strategy for the next question (maximum 2 sentences).]

Type of Question: [Based on your thought, choose one of the six Socratic question types: Definition, Hypothetical, Elenchus, Maieutic, Reductio, or Dialectic.]

Question: Formulate your question to challenge your opponent, aligned with the selected type.

Figure 9: Prompt for asking the Socratic questions.

You are a financial expert participating in a structured Socratic debate on the short-term price movement of AAPL. Your objective is to defend your prediction against the critical questioning of your opponent. You will be provided with:

1. Your Original Arguments: Your initial stance and prediction regarding AAPL's short-term stock price movement.
2. Opponent's Rebuttal: A critique of your original argument, highlighting weaknesses or flaws.
3. Opponent's Arguments: Your opponent's prediction on AAPL's short-term stock price movement, predicting the opposite of your stance.
4. Questioning History: A transcript of all questions and answers exchanged during the debate rounds.

Task Description:

Respond to your opponent's most recent question with a direct and concise answer. Avoid deflecting or changing the subject.

Your goal is to address the question head-on, maintaining clarity and argumentative discipline.

Output Format:

Answer: [Provide a direct response, clearly addressing your opponent's question.]

Figure 10: Prompt for answering the Socratic questions.

You are acting as a financial judge tasked with evaluating the reasoning of two debaters (Expert A & Expert B) regarding AAPL's short-term stock price movement. You will assess their arguments, rebuttals, and cross-questioning based on the following information:

1. Expert A's Argument – Expert A's position and supporting reasoning about AAPL's stock price movement.
2. Expert B's Argument – Expert B's position and supporting reasoning, predicting the opposite movement.
3. Expert A's Rebuttal – Expert A's critique of Expert B's argument, identifying weaknesses or flaws.
4. Expert B's Rebuttal – Expert B's critique of Expert A's argument, identifying weaknesses or flaws.
5. Cross-questioning – A transcript of all questions and answers exchanged during the debate rounds.

Evaluate which debater presents the most reasonable and well-supported position. You should consider the following aspects in your evaluation:

1. Argument Strength: Which debater presents a more compelling argument, supported by stronger evidence or reasoning?
2. Effectiveness in Cross-questioning: Did the debater answer questions directly, or did they avoid or deflect them? Was the questioning stage used effectively to expose flaws in the opponent's argument?
3. Overall Persuasiveness: Taking into account the arguments, rebuttals, questioning, and conclusions, which debater made a more rational and well-supported case?
4. Clarity and Coherence: How clear, structured, and logically consistent were each debater's arguments and rebuttals?

Final Assessment Requirements:

1. Reasoning Process: Provide a clear and concise explanation (maximum two paragraphs) justifying your decision. Focus on the overall coherence, logical soundness, and evidential strength of each debater's arguments. Pay particular attention to how each debater handled the questioning stage — did they effectively respond to the questions, defend their own position under pressure, or expose weaknesses in their opponent's case? Highlight key contradictions, evasions, or strong rebuttals that influenced your decision-making.
2. Final Judgement: Based on your reasoning, determine which expert presented the stronger case and is more likely to be correct in predicting AAPL's short-term stock price movement.

Output Format:

Reasoning Process: Provide a clear and concise explanation (maximum two paragraphs) justifying your evaluation. Focus on:

- The overall coherence and logical soundness of each debater's argument.
- The effectiveness of their responses to cross-questioning and how they defended their position.
- Key contradictions or evasions in the opponent's argument that influenced your decision.

Final Judgement: Only answer either "Buy" , "Sell" , or "Hold" based on your reasoning. Do not give any reasons here

Figure 11: Prompt for making the final trading decision.

You are an expert financial analyst tasked with analyzing a debate conversation between a bullish agent, a bearish agent, and a trader.

Given the following debate conversation, identify which Socratic question the trader emphasizes most when making the final trading decision.

Debate Conversation:

{debate_conversation}

Trader's Final Decision: {final_decision}

There are six types of Socratic questions:

1. Definition: Clarifies ambiguous terms or metrics.
2. Hypothetical: Tests reasoning under alternative scenarios.
3. Elenchus: Exposes internal contradictions in reasoning.
4. Maieutic: Uncovers unspoken assumptions or beliefs.
5. Reductio: Pushes logic to an extreme or absurd outcome.
6. Dialectic: Synthesizes opposing views to find deeper truth.

Instructions:

- Carefully read the entire debate conversation.
- Identify which question the trader references or emphasizes most in their final trading decision.
- Select only ONE question type from the list above.
- Provide a brief justification for your choice.

Output Format: Question Type: [Definition / Hypothetical / Elenchus / Maieutic / Reductio / Dialectic]

Justification: [One sentence explaining why this question influenced the trader's decision most.]

Figure 12: Prompt for identifying the most influential Socratic question type for trading decisions.

Before beginning this annotation task, please read the following information. By continuing, you acknowledge that the labels you provide will be used for research purposes, and that your participation is voluntary.

Use of Data:

- Your annotations will be collected and analyzed solely for academic research purposes.
- The data may be shared as an anonymized dataset for research under a non-commercial license.
- All collected data will remain anonymous and confidential.

Acknowledgment:

By proceeding to the annotation task, you acknowledge that you have read this information and agree that your annotations may be used for research purposes. Please contact [AUTHOR'S EMAIL] if you have any questions regarding this study. If you do not wish to participate, you may exit the task at any time.

Figure 13: Annotator Information and Consent Notice

Task Overview:

You are asked to analyze a debate conversation between a bullish agent, a bearish agent, and a trader. Your goal is to identify which type of Socratic question the trader emphasizes most when making their final trading decision.

Materials Provided:

1. Debate Conversation:

debate_conversation

2. Trader's Final Decision:

final_decision

There are six types of Socratic questions:

1. Definition: Clarifies ambiguous terms or metrics.
2. Hypothetical: Tests reasoning under alternative scenarios.
3. Elenchus: Exposes internal contradictions in reasoning.
4. Maieutic: Uncovers unspoken assumptions or beliefs.
5. Reductio: Pushes logic to an extreme or absurd outcome.
6. Dialectic: Synthesizes opposing views to find deeper truth.

Instructions:

- Carefully read the entire debate conversation.
- Determine which Socratic question type the trader references or emphasizes most in their final decision.
- Select only one question type from the list above.
- Provide a brief justification (one sentence) explaining why you chose this question type and how it influenced the trader's decision.

Figure 14: Annotation Instructions.

	NVDA			AAPL		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	47.24	0.99	36.88	22.00	0.77	33.36
MACD	7.21	0.32	41.33	-2.32	-0.15	20.89
PEN	-15.15 $\pm$ 5.7	-0.12 $\pm$ 0.4	59.00 $\pm$ 4.8	-8.14 $\pm$ 9.3	-0.23 $\pm$ 0.4	35.47 $\pm$ 1.6
FININ	-11.60 $\pm$ 5.4	-0.04 $\pm$ 0.3	63.41 $\pm$ 5.8	-1.65 $\pm$ 9.1	0.01 $\pm$ 0.3	39.05 $\pm$ 6.2
LLMFactor	6.71 $\pm$ 11.5	0.34 $\pm$ 0.2	48.43 $\pm$ 6.0	11.44 $\pm$ 11.5	0.45 $\pm$ 0.2	25.91 $\pm$ 6.5
FinMem	40.67 $\pm$ 10.3	0.90 $\pm$ 0.2	42.74 $\pm$ 2.7	15.59 $\pm$ 10.2	0.58 $\pm$ 0.1	22.74 $\pm$ 4.1
FinAgent	45.00 $\pm$ 7.5	0.95 $\pm$ 0.1	42.73 $\pm$ 6.2	36.46 $\pm$ 10.8	1.16 $\pm$ 0.1	31.54 $\pm$ 3.4
FinCon	71.43 $\pm$ 13.7	1.29 $\pm$ 0.2	43.48 $\pm$ 2.5	49.16 $\pm$ 15.8	1.48 $\pm$ 0.2	15.61 $\pm$ 2.5
TA	85.48 $\pm$ 13.1	1.45 $\pm$ 0.3	48.33 $\pm$ 3.3	50.59 $\pm$ 6.9	1.52 $\pm$ 0.3	21.94 $\pm$ 4.0
TiMi	91.31 $\pm$ 12.4	1.51 $\pm$ 0.1	39.42 $\pm$ 1.0	59.07 $\pm$ 11.6	1.93 $\pm$ 0.2	18.25 $\pm$ 1.2
FinRAQ (GP)	129.53 $\pm$ 4.5	1.89 $\pm$ 0.1	31.84 $\pm$ 3.4	86.19 $\pm$ 3.6	2.27 $\pm$ 0.1	9.35 $\pm$ 3.8
FinRAQ (CL)	129.43 $\pm$ 4.0	1.89 $\pm$ 0.2	34.59 $\pm$ 2.6	89.44 $\pm$ 0.6	2.34 $\pm$ 0.1	11.82 $\pm$ 0.3
FinRAQ (GE)	130.10 $\pm$ 2.4	1.90 $\pm$ 0.1	30.62 $\pm$ 0.7	87.68 $\pm$ 3.8	2.30 $\pm$ 0.0	9.15 $\pm$ 0.4
FinRAQ (LL)	118.98 $\pm$ 3.2	1.79 $\pm$ 0.1	35.77 $\pm$ 2.2	79.46 $\pm$ 3.4	2.14 $\pm$ 0.0	10.69 $\pm$ 0.6
	CIFR*			QBTS*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	126.15	1.25	71.74	233.61	1.50	65.96
MACD	22.44	0.62	70.52	236.79	1.55	59.86
PEN	25.95 $\pm$ 11.3	0.76 $\pm$ 0.3	79.54 $\pm$ 5.5	57.81 $\pm$ 11.4	1.02 $\pm$ 0.3	81.35 $\pm$ 5.9
FININ	34.00 $\pm$ 11.8	0.82 $\pm$ 0.2	72.99 $\pm$ 4.8	58.19 $\pm$ 8.2	1.04 $\pm$ 0.2	74.21 $\pm$ 3.6
LLMFactor	52.36 $\pm$ 11.1	0.93 $\pm$ 0.3	82.34 $\pm$ 4.0	91.39 $\pm$ 0.3	1.19 $\pm$ 0.3	67.69 $\pm$ 5.3
FinMem	61.88 $\pm$ 10.1	0.97 $\pm$ 0.1	73.46 $\pm$ 4.3	105.83 $\pm$ 8.9	1.20 $\pm$ 0.1	59.77 $\pm$ 2.5
FinAgent	125.11 $\pm$ 10.9	1.28 $\pm$ 0.3	73.82 $\pm$ 6.3	139.09 $\pm$ 11.7	1.28 $\pm$ 0.3	51.77 $\pm$ 4.5
FinCon	150.04 $\pm$ 9.6	1.37 $\pm$ 0.2	77.16 $\pm$ 5.5	189.35 $\pm$ 8.3	1.42 $\pm$ 0.1	51.31 $\pm$ 1.2
TA	197.28 $\pm$ 14.4	1.52 $\pm$ 0.2	83.44 $\pm$ 1.1	284.03 $\pm$ 7.7	1.68 $\pm$ 0.1	54.02 $\pm$ 3.4
TiMi	202.51 $\pm$ 10.3	1.53 $\pm$ 0.0	72.09 $\pm$ 2.7	274.71 $\pm$ 10.4	1.68 $\pm$ 0.0	56.55 $\pm$ 0.3
FinRAQ (GP)	249.69 $\pm$ 2.7	1.67 $\pm$ 0.1	57.14 $\pm$ 5.0	346.06 $\pm$ 3.6	1.81 $\pm$ 0.0	50.66 $\pm$ 3.8
FinRAQ (CL)	248.78 $\pm$ 4.7	1.65 $\pm$ 0.1	55.47 $\pm$ 4.0	345.16 $\pm$ 6.3	1.78 $\pm$ 0.1	51.69 $\pm$ 1.7
FinRAQ (GE)	254.11 $\pm$ 5.1	1.68 $\pm$ 0.1	53.44 $\pm$ 4.8	358.48 $\pm$ 0.7	1.87 $\pm$ 0.0	46.63 $\pm$ 2.5
FinRAQ (LL)	226.80 $\pm$ 1.7	1.60 $\pm$ 0.0	70.38 $\pm$ 5.7	320.95 $\pm$ 7.7	1.76 $\pm$ 0.1	55.20 $\pm$ 0.9

Table 11: Performance comparison on technology sector stocks ($p < 0.05$). "*" denotes mid-cap stocks from the Russell 2000 index. Bold values indicate the best performance. GP, CL, GE, and LL denote FinRAQ with GPT-5.4, Claude-Sonnet-4.6, Gemini-3-Pro, and Llama-4-Maverick-17B-128E-Instruct backbones, respectively.

	JPM			MS		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	30.34	1.10	24.42	42.94	1.28	29.24
MACD	13.59	0.72	18.68	14.37	0.74	17.03
PEN	4.30 $\pm$ 11.0	0.21 $\pm$ 0.1	27.14 $\pm$ 3.6	1.95 $\pm$ 11.2	0.15 $\pm$ 0.1	37.63 $\pm$ 4.4
FININ	7.62 $\pm$ 5.5	0.34 $\pm$ 0.3	26.68 $\pm$ 6.0	6.24 $\pm$ 10.2	0.28 $\pm$ 0.3	34.99 $\pm$ 6.0
LLMFactor	37.01 $\pm$ 5.3	1.30 $\pm$ 0.2	19.83 $\pm$ 4.2	25.10 $\pm$ 4.2	0.84 $\pm$ 0.3	27.42 $\pm$ 4.3
FinMem	52.37 $\pm$ 10.3	1.73 $\pm$ 0.2	16.19 $\pm$ 0.6	24.34 $\pm$ 9.7	0.81 $\pm$ 0.2	21.74 $\pm$ 3.4
FinAgent	54.71 $\pm$ 9.5	1.79 $\pm$ 0.1	18.26 $\pm$ 1.5	73.85 $\pm$ 9.1	1.95 $\pm$ 0.1	18.69 $\pm$ 2.6
FinCon	71.76 $\pm$ 10.9	2.21 $\pm$ 0.1	16.01 $\pm$ 1.2	73.88 $\pm$ 11.0	1.95 $\pm$ 0.2	38.12 $\pm$ 5.3
TA	92.35 $\pm$ 7.9	2.67 $\pm$ 0.2	22.32 $\pm$ 5.1	98.28 $\pm$ 14.2	2.40 $\pm$ 0.2	23.96 $\pm$ 4.0
TiMi	80.38 $\pm$ 12.5	2.41 $\pm$ 0.2	16.32 $\pm$ 3.4	104.62 $\pm$ 13.8	2.51 $\pm$ 0.1	17.85 $\pm$ 4.2
FinRAQ (GP)	114.95 $\pm$ 2.2	3.13 $\pm$ 0.1	9.43 $\pm$ 4.1	126.98 $\pm$ 3.1	2.87 $\pm$ 0.0	15.02 $\pm$ 1.3
FinRAQ (CL)	116.39 $\pm$ 1.3	3.16 $\pm$ 0.2	11.62 $\pm$ 1.1	129.52 $\pm$ 3.6	2.91 $\pm$ 0.1	12.04 $\pm$ 0.3
FinRAQ (GE)	115.41 $\pm$ 2.4	3.14 $\pm$ 0.2	8.80 $\pm$ 0.5	131.14 $\pm$ 4.7	2.94 $\pm$ 0.1	13.42 $\pm$ 0.3
FinRAQ (LL)	108.86 $\pm$ 5.6	3.01 $\pm$ 0.1	13.93 $\pm$ 0.2	114.69 $\pm$ 5.3	2.68 $\pm$ 0.2	16.41 $\pm$ 0.9
	MARA*			WULF*		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-42.71	-0.19	75.46	183.62	1.43	74.60
MACD	-38.37	-0.44	70.44	-44.49	-0.44	84.49
PEN	-13.96 $\pm$ 6.4	0.27 $\pm$ 0.2	86.35 $\pm$ 2.3	0.53 $\pm$ 5.1	0.57 $\pm$ 0.3	69.29 $\pm$ 4.3
FININ	-2.95 $\pm$ 10.1	0.40 $\pm$ 0.3	54.54 $\pm$ 4.4	26.93 $\pm$ 5.3	0.76 $\pm$ 0.2	78.72 $\pm$ 5.7
LLMFactor	-1.42 $\pm$ 11.3	0.41 $\pm$ 0.2	63.70 $\pm$ 2.0	56.81 $\pm$ 12.2	0.93 $\pm$ 0.2	69.82 $\pm$ 2.0
FinMem	3.48 $\pm$ 11.6	0.46 $\pm$ 0.3	54.59 $\pm$ 4.0	82.51 $\pm$ 11.7	1.13 $\pm$ 0.3	81.81 $\pm$ 3.4
FinAgent	10.89 $\pm$ 8.4	0.54 $\pm$ 0.3	71.79 $\pm$ 3.5	161.72 $\pm$ 3.5	1.36 $\pm$ 0.3	75.67 $\pm$ 5.7
FinCon	18.56 $\pm$ 7.0	0.61 $\pm$ 0.3	73.13 $\pm$ 0.3	127.56 $\pm$ 14.4	1.26 $\pm$ 0.3	76.64 $\pm$ 6.5
TA	35.75 $\pm$ 12.7	0.78 $\pm$ 0.1	89.46 $\pm$ 5.4	215.54 $\pm$ 13.3	1.54 $\pm$ 0.2	61.88 $\pm$ 1.8
TiMi	37.34 $\pm$ 12.7	0.79 $\pm$ 0.2	60.73 $\pm$ 3.3	217.68 $\pm$ 13.9	1.62 $\pm$ 0.1	73.12 $\pm$ 3.7
FinRAQ (GP)	56.68 $\pm$ 4.8	0.93 $\pm$ 0.1	42.90 $\pm$ 1.6	255.15 $\pm$ 3.8	1.65 $\pm$ 0.1	50.17 $\pm$ 4.3
FinRAQ (CL)	58.92 $\pm$ 0.7	0.95 $\pm$ 0.1	39.44 $\pm$ 0.9	253.51 $\pm$ 3.6	1.62 $\pm$ 0.1	52.88 $\pm$ 4.3
FinRAQ (GE)	59.42 $\pm$ 3.9	0.96 $\pm$ 0.1	38.73 $\pm$ 0.3	267.67 $\pm$ 5.0	1.74 $\pm$ 0.1	43.24 $\pm$ 1.4
FinRAQ (LL)	51.47 $\pm$ 2.4	0.90 $\pm$ 0.0	48.69 $\pm$ 3.9	235.91 $\pm$ 6.4	1.67 $\pm$ 0.1	67.42 $\pm$ 1.6

Table 12: Performance comparison on financial sector stocks ($p < 0.05$). "*" denotes mid-cap stocks from the Russell 2000 index. Bold values indicate the best performance. GP, CL, GE, and LL denote FinRAQ with GPT-5.4, Claude-Sonnet-4.6, Gemini-3-Pro, and Llama-4-Maverick-17B-128E-Instruct backbones, respectively.

	JNJ			UNH		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	27.75	1.35	14.45	-21.19	-0.40	61.39
MACD	22.17	1.41	9.32	-5.62	-0.08	29.58
PEN	3.26 $\pm$ 8.2	0.16 $\pm$ 0.3	19.39 $\pm$ 2.0	-19.16 $\pm$ 7.5	-0.35 $\pm$ 0.3	68.06 $\pm$ 6.4
FININ	7.61 $\pm$ 6.7	0.39 $\pm$ 0.3	27.00 $\pm$ 3.9	-17.00 $\pm$ 8.2	-0.29 $\pm$ 0.3	43.33 $\pm$ 5.0
LLMFactor	14.48 $\pm$ 7.7	0.73 $\pm$ 0.1	15.89 $\pm$ 3.1	-1.44 $\pm$ 0.5	0.12 $\pm$ 0.2	36.73 $\pm$ 5.5
FinMem	24.08 $\pm$ 11.5	1.19 $\pm$ 0.2	11.78 $\pm$ 2.2	-0.38 $\pm$ 8.0	0.16 $\pm$ 0.2	56.12 $\pm$ 6.6
FinAgent	40.15 $\pm$ 13.0	1.88 $\pm$ 0.1	21.22 $\pm$ 0.5	1.86 $\pm$ 0.5	0.20 $\pm$ 0.2	53.67 $\pm$ 5.2
FinCon	52.35 $\pm$ 11.0	2.35 $\pm$ 0.2	10.07 $\pm$ 2.3	25.90 $\pm$ 6.4	0.72 $\pm$ 0.1	31.23 $\pm$ 3.8
TA	63.78 $\pm$ 14.7	2.77 $\pm$ 0.1	15.85 $\pm$ 3.5	58.81 $\pm$ 3.5	1.28 $\pm$ 0.1	29.77 $\pm$ 2.6
TiMi	61.71 $\pm$ 11.6	2.70 $\pm$ 0.1	12.58 $\pm$ 2.2	48.95 $\pm$ 9.7	1.12 $\pm$ 0.2	25.42 $\pm$ 4.0
FinRAQ (GP)	89.51 $\pm$ 4.2	3.64 $\pm$ 0.0	9.12 $\pm$ 0.2	78.89 $\pm$ 6.2	1.57 $\pm$ 0.0	20.84 $\pm$ 0.4
FinRAQ (CL)	88.18 $\pm$ 2.5	3.60 $\pm$ 0.0	10.63 $\pm$ 0.1	80.47 $\pm$ 2.6	1.58 $\pm$ 0.0	23.53 $\pm$ 1.0
FinRAQ (GE)	91.03 $\pm$ 2.3	3.69 $\pm$ 0.1	7.98 $\pm$ 0.1	80.80 $\pm$ 2.3	1.59 $\pm$ 0.1	20.33 $\pm$ 0.6
FinRAQ (LL)	80.80 $\pm$ 2.2	3.36 $\pm$ 0.1	11.50 $\pm$ 0.7	71.63 $\pm$ 2.4	1.47 $\pm$ 0.0	24.14 $\pm$ 1.0
	IOVA*			IBRX*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	-52.33	-0.35	90.33	44.52	0.83	79.34
MACD	-51.26	-0.61	85.42	14.26	0.51	65.24
PEN	-12.15 $\pm$ 7.1	0.29 $\pm$ 0.3	89.50 $\pm$ 3.7	1.49 $\pm$ 12.1	0.52 $\pm$ 0.3	82.37 $\pm$ 7.9
FININ	-16.74 $\pm$ 8.5	0.23 $\pm$ 0.3	62.64 $\pm$ 3.2	-9.28 $\pm$ 10.1	0.43 $\pm$ 0.3	76.35 $\pm$ 4.6
LLMFactor	-4.79 $\pm$ 8.3	0.37 $\pm$ 0.2	72.78 $\pm$ 3.4	20.33 $\pm$ 0.2	0.72 $\pm$ 0.2	85.15 $\pm$ 2.5
FinMem	19.71 $\pm$ 8.2	0.62 $\pm$ 0.2	57.97 $\pm$ 1.8	32.84 $\pm$ 6.0	0.79 $\pm$ 0.2	90.94 $\pm$ 6.5
FinAgent	28.40 $\pm$ 7.9	0.70 $\pm$ 0.1	71.77 $\pm$ 2.0	54.48 $\pm$ 5.1	0.93 $\pm$ 0.1	80.24 $\pm$ 5.8
FinCon	40.35 $\pm$ 11.3	0.84 $\pm$ 0.1	76.28 $\pm$ 1.0	60.83 $\pm$ 5.8	0.96 $\pm$ 0.2	61.41 $\pm$ 3.5
TA	68.31 $\pm$ 11.4	0.98 $\pm$ 0.1	53.78 $\pm$ 5.5	103.49 $\pm$ 15.5	1.17 $\pm$ 0.2	59.48 $\pm$ 6.2
TiMi	62.48 $\pm$ 8.7	0.97 $\pm$ 0.2	59.27 $\pm$ 4.6	109.93 $\pm$ 8.7	1.20 $\pm$ 0.1	58.57 $\pm$ 3.8
FinRAQ (GP)	77.21 $\pm$ 1.5	1.10 $\pm$ 0.0	50.77 $\pm$ 3.6	124.27 $\pm$ 1.5	1.32 $\pm$ 0.0	53.76 $\pm$ 4.0
FinRAQ (CL)	77.80 $\pm$ 1.8	1.08 $\pm$ 0.1	49.39 $\pm$ 1.8	124.73 $\pm$ 1.6	1.31 $\pm$ 0.0	51.78 $\pm$ 4.0
FinRAQ (GE)	80.89 $\pm$ 6.4	1.11 $\pm$ 0.1	49.19 $\pm$ 2.2	129.47 $\pm$ 2.9	1.39 $\pm$ 0.1	44.60 $\pm$ 1.3
FinRAQ (LL)	73.20 $\pm$ 1.5	1.08 $\pm$ 0.1	57.78 $\pm$ 2.9	114.83 $\pm$ 4.6	1.26 $\pm$ 0.1	57.97 $\pm$ 3.5

Table 13: Performance comparison on healthcare sector stocks ($p < 0.05$). "*" denotes mid-cap stocks from the Russell 2000 index. Bold values indicate the best performance. GP, CL, GE, and LL denote FinRAQ with GPT-5.4, Claude-Sonnet-4.6, Gemini-3-Pro, and Llama-4-Maverick-17B-128E-Instruct backbones, respectively.

	XOM			CVX		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	24.34	1.00	18.92	15.56	0.67	20.64
MACD	7.19	0.39	19.27	2.15	0.09	21.85
PEN	8.91 $\pm$ 11.5	0.40 $\pm$ 0.2	28.23 $\pm$ 6.6	7.56 $\pm$ 7.7	0.35 $\pm$ 0.2	30.77 $\pm$ 5.8
FININ	0.67 $\pm$ 8.0	0.05 $\pm$ 0.3	33.06 $\pm$ 5.6	13.25 $\pm$ 10.3	0.58 $\pm$ 0.2	16.10 $\pm$ 1.7
LLMFactor	24.55 $\pm$ 9.3	1.00 $\pm$ 0.2	23.12 $\pm$ 5.5	23.20 $\pm$ 8.7	0.96 $\pm$ 0.2	17.57 $\pm$ 2.3
FinMem	38.03 $\pm$ 11.1	1.47 $\pm$ 0.2	22.08 $\pm$ 1.5	36.62 $\pm$ 12.0	1.42 $\pm$ 0.1	16.73 $\pm$ 1.6
FinAgent	57.75 $\pm$ 12.6	2.07 $\pm$ 0.2	15.20 $\pm$ 1.8	40.06 $\pm$ 12.6	1.54 $\pm$ 0.2	15.46 $\pm$ 3.4
FinCon	62.08 $\pm$ 6.9	2.20 $\pm$ 0.1	11.72 $\pm$ 4.1	45.48 $\pm$ 10.5	1.71 $\pm$ 0.1	22.73 $\pm$ 5.8
TA	70.22 $\pm$ 10.0	2.42 $\pm$ 0.2	11.44 $\pm$ 2.4	53.24 $\pm$ 9.0	1.95 $\pm$ 0.2	20.81 $\pm$ 4.1
TiMi	70.51 $\pm$ 8.3	2.50 $\pm$ 0.1	12.10 $\pm$ 2.7	52.20 $\pm$ 9.0	1.91 $\pm$ 0.1	19.38 $\pm$ 1.0
FinRAQ (GP)	94.78 $\pm$ 3.6	3.05 $\pm$ 0.0	9.24 $\pm$ 0.5	71.68 $\pm$ 0.6	2.47 $\pm$ 0.0	14.98 $\pm$ 1.9
FinRAQ (CL)	94.54 $\pm$ 0.1	3.05 $\pm$ 0.1	9.38 $\pm$ 0.8	72.50 $\pm$ 3.0	2.49 $\pm$ 0.1	14.46 $\pm$ 1.2
FinRAQ (GE)	97.21 $\pm$ 3.6	3.11 $\pm$ 0.1	9.18 $\pm$ 0.6	73.95 $\pm$ 5.6	2.53 $\pm$ 0.1	12.30 $\pm$ 0.6
FinRAQ (LL)	85.58 $\pm$ 1.1	2.83 $\pm$ 0.1	11.63 $\pm$ 1.2	66.86 $\pm$ 1.7	2.34 $\pm$ 0.1	18.79 $\pm$ 0.5
	RIG*			PTEN*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	12.94	0.47	68.11	-11.58	-0.03	55.90
MACD	26.07	0.72	39.43	-29.14	-0.83	62.93
PEN	-3.75 $\pm$ 8.1	0.18 $\pm$ 0.4	48.48 $\pm$ 1.7	-15.83 $\pm$ 8.1	-0.12 $\pm$ 0.2	64.24 $\pm$ 6.1
FININ	1.57 $\pm$ 7.0	0.28 $\pm$ 7.0	61.03 $\pm$ 2.9	-7.28 $\pm$ 6.0	0.07 $\pm$ 0.1	60.50 $\pm$ 2.2
LLMFactor	14.30 $\pm$ 11.2	0.48 $\pm$ 0.2	64.59 $\pm$ 2.0	6.37 $\pm$ 11.5	0.34 $\pm$ 0.2	61.48 $\pm$ 11.2
FinMem	42.37 $\pm$ 10.4	0.87 $\pm$ 0.2	55.34 $\pm$ 2.6	10.67 $\pm$ 9.8	0.42 $\pm$ 0.2	62.60 $\pm$ 4.5
FinAgent	50.29 $\pm$ 13.1	0.97 $\pm$ 0.3	40.18 $\pm$ 0.5	39.35 $\pm$ 11.3	0.87 $\pm$ 0.3	58.10 $\pm$ 0.3
FinCon	54.09 $\pm$ 7.9	1.01 $\pm$ 0.1	52.67 $\pm$ 1.3	43.54 $\pm$ 0.4	0.93 $\pm$ 0.1	57.88 $\pm$ 4.9
TA	68.68 $\pm$ 10.3	1.17 $\pm$ 0.1	34.35 $\pm$ 0.5	54.66 $\pm$ 14.4	1.21 $\pm$ 0.1	60.52 $\pm$ 5.8
TiMi	77.61 $\pm$ 9.8	1.27 $\pm$ 0.1	59.30 $\pm$ 4.2	53.69 $\pm$ 12.3	1.19 $\pm$ 0.2	58.34 $\pm$ 0.2
FinRAQ (GP)	97.15 $\pm$ 4.5	1.44 $\pm$ 0.1	28.70 $\pm$ 0.2	71.37 $\pm$ 2.1	1.29 $\pm$ 0.1	34.45 $\pm$ 2.4
FinRAQ (CL)	99.25 $\pm$ 1.3	1.46 $\pm$ 0.0	27.72 $\pm$ 2.4	72.20 $\pm$ 6.3	1.29 $\pm$ 0.0	33.10 $\pm$ 0.0
FinRAQ (GE)	101.30 $\pm$ 4.1	1.49 $\pm$ 0.0	21.12 $\pm$ 1.5	71.92 $\pm$ 0.6	1.30 $\pm$ 0.1	24.22 $\pm$ 2.4
FinRAQ (LL)	88.74 $\pm$ 7.5	1.37 $\pm$ 0.0	32.50 $\pm$ 0.5	65.19 $\pm$ 0.8	1.21 $\pm$ 0.0	44.05 $\pm$ 1.4

Table 14: Performance comparison on energy sector stocks ($p < 0.05$). "*" denotes mid-cap stocks from the Russell 2000 index. Bold values indicate the best performance. GP, CL, GE, and LL denote FinRAQ with GPT-5.4, Claude-Sonnet-4.6, Gemini-3-Pro, and Llama-4-Maverick-17B-128E-Instruct backbones, respectively.

	WMT			KO		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	49.63	1.80	21.93	20.74	1.15	15.50
MACD	20.58	1.16	11.11	10.86	0.75	15.23
PEN	2.24 $\pm$ 7.1	0.13 $\pm$ 0.2	20.32 $\pm$ 5.2	6.07 $\pm$ 0.2	0.33 $\pm$ 5.2	15.25 $\pm$ 5.2
FININ	15.89 $\pm$ 12.3	0.67 $\pm$ 0.2	24.69 $\pm$ 1.5	6.39 $\pm$ 9.1	0.34 $\pm$ 0.1	15.73 $\pm$ 4.4
LLMFactor	48.73 $\pm$ 7.0	1.78 $\pm$ 0.3	13.09 $\pm$ 0.2	10.15 $\pm$ 2.3	0.57 $\pm$ 0.2	14.83 $\pm$ 2.3
FinMem	52.68 $\pm$ 7.5	1.89 $\pm$ 0.1	13.77 $\pm$ 2.3	21.18 $\pm$ 2.3	1.17 $\pm$ 0.2	17.31 $\pm$ 4.2
FinAgent	83.72 $\pm$ 17.8	2.73 $\pm$ 0.1	16.06 $\pm$ 9.7	33.98 $\pm$ 9.7	1.82 $\pm$ 0.1	12.02 $\pm$ 3.8
FinCon	92.40 $\pm$ 11.4	2.94 $\pm$ 0.1	10.05 $\pm$ 0.3	57.64 $\pm$ 7.6	2.88 $\pm$ 0.2	9.47 $\pm$ 0.6
TA	121.01 $\pm$ 10.3	3.59 $\pm$ 0.2	12.77 $\pm$ 4.0	59.28 $\pm$ 10.9	2.95 $\pm$ 0.1	8.79 $\pm$ 3.7
TiMi	112.84 $\pm$ 13.4	3.41 $\pm$ 0.1	18.62 $\pm$ 2.6	55.44 $\pm$ 10.0	2.21 $\pm$ 0.1	9.01 $\pm$ 0.5
FinRAQ (GP)	140.92 $\pm$ 2.6	4.00 $\pm$ 0.0	8.64 $\pm$ 0.2	81.95 $\pm$ 4.8	3.86 $\pm$ 0.1	6.86 $\pm$ 1.3
FinRAQ (CL)	140.10 $\pm$ 3.1	3.99 $\pm$ 0.1	13.01 $\pm$ 0.9	84.11 $\pm$ 5.2	3.94 $\pm$ 0.1	4.99 $\pm$ 0.2
FinRAQ (GE)	142.32 $\pm$ 4.0	4.03 $\pm$ 0.1	8.19 $\pm$ 0.8	82.94 $\pm$ 4.6	3.90 $\pm$ 0.1	6.15 $\pm$ 0.1
FinRAQ (LL)	127.12 $\pm$ 3.9	3.72 $\pm$ 0.0	12.29 $\pm$ 0.5	77.83 $\pm$ 4.1	3.70 $\pm$ 0.1	7.86 $\pm$ 0.1
	LAUR*			LRN*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	55.32	1.56	14.65	17.65	0.60	64.07
MACD	14.04	0.66	17.92	18.08	0.65	57.71
PEN	1.18 $\pm$ 10.0	0.12 $\pm$ 0.1	44.18 $\pm$ 3.9	8.99 $\pm$ 10.6	0.48 $\pm$ 0.1	70.01 $\pm$ 3.3
FININ	1.55 $\pm$ 6.6	0.13 $\pm$ 0.3	39.26 $\pm$ 2.4	2.93 $\pm$ 9.5	0.41 $\pm$ 0.1	58.47 $\pm$ 9.5
LLMFactor	38.38 $\pm$ 13.0	1.17 $\pm$ 0.1	22.46 $\pm$ 0.4	36.11 $\pm$ 6.6	0.78 $\pm$ 0.2	40.51 $\pm$ 2.1
FinMem	56.91 $\pm$ 13.4	1.59 $\pm$ 0.1	36.36 $\pm$ 3.7	40.68 $\pm$ 9.5	0.90 $\pm$ 0.1	74.28 $\pm$ 5.0
FinAgent	96.57 $\pm$ 11.2	2.36 $\pm$ 0.2	17.03 $\pm$ 1.3	55.98 $\pm$ 11.8	0.97 $\pm$ 0.2	33.24 $\pm$ 1.6
FinCon	119.40 $\pm$ 12.0	2.74 $\pm$ 0.1	21.55 $\pm$ 1.2	52.51 $\pm$ 1.2	0.97 $\pm$ 0.2	45.10 $\pm$ 1.7
TA	154.38 $\pm$ 10.4	3.26 $\pm$ 0.2	14.93 $\pm$ 1.0	76.87 $\pm$ 10.8	1.29 $\pm$ 0.1	59.03 $\pm$ 2.8
TiMi	198.56 $\pm$ 11.5	3.84 $\pm$ 0.1	12.15 $\pm$ 0.3	65.74 $\pm$ 12.3	1.11 $\pm$ 0.1	43.75 $\pm$ 0.2
FinRAQ (GP)	229.92 $\pm$ 3.4	4.20 $\pm$ 0.1	9.13 $\pm$ 0.2	87.39 $\pm$ 5.1	1.39 $\pm$ 0.1	30.72 $\pm$ 2.9
FinRAQ (CL)	226.88 $\pm$ 4.4	4.17 $\pm$ 0.0	10.42 $\pm$ 0.4	87.70 $\pm$ 5.0	1.39 $\pm$ 0.0	28.78 $\pm$ 2.5
FinRAQ (GE)	236.18 $\pm$ 4.7	4.27 $\pm$ 0.1	8.16 $\pm$ 0.3	90.84 $\pm$ 0.3	1.45 $\pm$ 0.1	27.96 $\pm$ 1.2
FinRAQ (LL)	211.84 $\pm$ 5.5	4.00 $\pm$ 0.1	11.67 $\pm$ 0.1	82.95 $\pm$ 2.9	1.28 $\pm$ 0.1	39.11 $\pm$ 3.4

Table 15: Performance comparison on consumer staples sector stocks ($p < 0.05$). "*" denotes mid-cap stocks from the Russell 2000 index. Bold values indicate the best performance. GP, CL, GE, and LL denote FinRAQ with GPT-5.4, Claude-Sonnet-4.6, Gemini-3-Pro, and Llama-4-Maverick-17B-128E-Instruct backbones, respectively.

	BTC-USD		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	2.46	0.21	49.74
MACD	0.63	0.08	32.49
PEN	-17.37 $\pm$ 11.5	-0.18 $\pm$ 0.3	58.77 $\pm$ 2.3
FININ	-3.33 $\pm$ 7.1	0.14 $\pm$ 0.2	44.01 $\pm$ 1.3
LLMFactor	48.00 $\pm$ 15.8	1.00 $\pm$ 0.1	40.00 $\pm$ 5.4
FinMem	56.75 $\pm$ 9.8	1.12 $\pm$ 0.1	44.31 $\pm$ 4.3
FinAgent	79.24 $\pm$ 13.3	1.39 $\pm$ 0.2	37.16 $\pm$ 4.7
FinCon	81.12 $\pm$ 10.2	1.42 $\pm$ 0.1	36.30 $\pm$ 5.6
TA	94.07 $\pm$ 9.8	1.56 $\pm$ 0.2	28.15 $\pm$ 0.9
TiMi	102.19 $\pm$ 10.4	1.74 $\pm$ 0.1	21.78 $\pm$ 3.1
FinRAQ (GP)	128.36 $\pm$ 1.4	1.89 $\pm$ 0.1	15.97 $\pm$ 0.6
FinRAQ (CL)	134.37 $\pm$ 4.5	1.95 $\pm$ 0.0	13.11 $\pm$ 1.2
FinRAQ (GE)	129.31 $\pm$ 5.9	1.90 $\pm$ 0.1	14.20 $\pm$ 1.2
FinRAQ (LL)	118.42 $\pm$ 1.1	1.80 $\pm$ 0.0	20.14 $\pm$ 0.4

Table 16: Performance comparison on BTC-USD ($p < 0.05$). Bold values indicate the best performance. GP, CL, GE, and LL denote FinRAQ with GPT-5.4, Claude-Sonnet-4.6, Gemini-3-Pro, and Llama-4-Maverick-17B-128E-Instruct backbones, respectively.

Period	MARA* (Financial)			UNH (Healthcare)		
	2025-04-30 ~2025-10-21			2025-05-07 ~2025-09-21		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	53.74	0.95	49.77	79.47	1.77	16.44
MACD	-18.97	-0.09	25.82	59.33	1.77	10.49
PEN	8.12 $\pm$ 6.8	0.09 $\pm$ 0.2	63.68 $\pm$ 3.2	-1.21 $\pm$ 10.0	-0.43 $\pm$ 0.1	35.44 $\pm$ 5.0
FININ	-8.93 $\pm$ 11.3	-0.05 $\pm$ 0.2	67.73 $\pm$ 9.3	9.78 $\pm$ 9.6	0.38 $\pm$ 0.1	29.47 $\pm$ 4.2
LLMFactor	2.25 $\pm$ 11.1	0.3 $\pm$ 0.1	64.71 $\pm$ 5.1	35.77 $\pm$ 8.7	1.12 $\pm$ 0.1	29.73 $\pm$ 1.8
FinMem	28.76 $\pm$ 5.4	0.54 $\pm$ 0.2	59.44 $\pm$ 3.6	56.28 $\pm$ 9.3	1.38 $\pm$ 0.1	18.36 $\pm$ 1.1
FinAgent	39.85 $\pm$ 10.3	0.85 $\pm$ 0.1	55.24 $\pm$ 4.1	68.91 $\pm$ 6.2	1.59 $\pm$ 0.1	13.67 $\pm$ 0.9
FinCon	43.43 $\pm$ 6.5	1.13 $\pm$ 0.1	57.58 $\pm$ 5.5	60.51 $\pm$ 9.5	1.64 $\pm$ 0.1	16.30 $\pm$ 4.1
TA	64.34 $\pm$ 12.8	1.29 $\pm$ 0.2	36.93 $\pm$ 1.3	72.96 $\pm$ 14.5	1.74 $\pm$ 0.1	15.77 $\pm$ 4.5
TiMi	72.10 $\pm$ 14.2	1.46 $\pm$ 0.1	37.00 $\pm$ 3.6	50.06 $\pm$ 13.5	1.26 $\pm$ 0.1	16.50 $\pm$ 1.7
FinRAQ	86.91 $\pm$ 5.5	1.91 $\pm$ 0.1	29.43 $\pm$ 2.4	96.71 $\pm$ 4.7	1.90 $\pm$ 0.0	9.18 $\pm$ 0.6
Period	IOVA* (Healthcare)			CVX (Energy)		
	2025-08-11 ~2026-02-02			2025-08-22 ~2026-02-13		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
B&H	79.66	1.07	32.71	39.81	1.61	8.31
MACD	-53.01	-1.18	43.64	10.84	0.54	15.14
PEN	-16.37 $\pm$ 10.6	-0.81 $\pm$ 0.1	51.1 $\pm$ 3.3	3.15 $\pm$ 11.3	0.10 $\pm$ 0.2	20.11 $\pm$ 4.9
FININ	7.67 $\pm$ 10.2	0.18 $\pm$ 0.1	51.15 $\pm$ 2.2	3.36 $\pm$ 11.3	0.08 $\pm$ 0.2	16.23 $\pm$ 4.0
LLMFactor	55.33 $\pm$ 12.3	0.94 $\pm$ 0.1	50.11 $\pm$ 6.9	21.88 $\pm$ 8.1	0.95 $\pm$ 0.2	11.87 $\pm$ 8.1
FinMem	66.42 $\pm$ 9.8	1.03 $\pm$ 0.1	53.29 $\pm$ 7.7	27.18 $\pm$ 10.2	1.03 $\pm$ 0.1	17.68 $\pm$ 3.4
FinAgent	70.91 $\pm$ 17.0	1.22 $\pm$ 0.1	46.48 $\pm$ 1.0	25.95 $\pm$ 10.9	1.28 $\pm$ 0.1	14.08 $\pm$ 1.2
FinCon	87.69 $\pm$ 9.6	1.74 $\pm$ 0.1	41.26 $\pm$ 4.2	20.97 $\pm$ 11.6	1.10 $\pm$ 0.1	18.73 $\pm$ 11.6
TA	131.62 $\pm$ 10.9	1.97 $\pm$ 0.1	26.63 $\pm$ 1.4	31.33 $\pm$ 6.8	1.27 $\pm$ 0.1	13.10 $\pm$ 1.8
TiMi	127.47 $\pm$ 16.1	1.93 $\pm$ 0.2	28.58 $\pm$ 2.6	30.65 $\pm$ 9.5	1.26 $\pm$ 0.1	10.76 $\pm$ 1.7
FinRAQ	140.34 $\pm$ 2.1	2.17 $\pm$ 0.2	22.06 $\pm$ 0.0	44.18 $\pm$ 1.0	1.75 $\pm$ 0.1	7.20 $\pm$ 0.3

Table 17: Performance comparison under bull market conditions (B&H ARR > 20%; $p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

Period	RIG* (Energy)			PTEN* (Energy)		
	2025-06-11 ~2025-12-02			2025-08-22 ~2026-02-13		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	85.93	1.35	23.19	111.65	1.78	19.97
MACD	-6.67	0.01	23.22	71.54	1.52	11.99
PEN	-14.4 $\pm$ 7.4	-0.03 $\pm$ 0.1	32.56 $\pm$ 5.3	-55.76 $\pm$ 12.1	-1.47 $\pm$ 0.1	40.12 $\pm$ 6.2
FININ	-20.57 $\pm$ 7.7	-1.00 $\pm$ 0.1	35.49 $\pm$ 3.9	-64.68 $\pm$ 9.4	-1.89 $\pm$ 0.1	42.01 $\pm$ 5.5
LLMFactor	8.43 $\pm$ 6.7	0.39 $\pm$ 0.2	27.79 $\pm$ 2.2	-4.04 $\pm$ 7.2	0.12 $\pm$ 0.2	28.72 $\pm$ 5.6
FinMem	18.08 $\pm$ 5.9	0.57 $\pm$ 0.1	27.19 $\pm$ 3.0	27.57 $\pm$ 12.4	0.70 $\pm$ 0.2	29.45 $\pm$ 2.9
FinAgent	66.06 $\pm$ 6.8	1.13 $\pm$ 0.1	26.93 $\pm$ 2.4	53.82 $\pm$ 10.5	1.40 $\pm$ 0.1	15.97 $\pm$ 5.2
FinCon	82.02 $\pm$ 15.0	1.26 $\pm$ 0.2	26.19 $\pm$ 0.6	140.16 $\pm$ 13.8	2.01 $\pm$ 0.1	14.34 $\pm$ 2.1
TA	95.48 $\pm$ 15.2	1.43 $\pm$ 0.1	22.49 $\pm$ 3.7	56.68 $\pm$ 9.0	1.13 $\pm$ 0.2	12.71 $\pm$ 4.6
TiMi	91.65 $\pm$ 10.2	1.38 $\pm$ 0.1	23.19 $\pm$ 1.3	85.91 $\pm$ 6.5	1.41 $\pm$ 0.2	10.25 $\pm$ 1.2
FinRAQ	108.89 $\pm$ 4.1	1.55 $\pm$ 0.0	21.63 $\pm$ 0.5	168.19 $\pm$ 12.6	2.27 $\pm$ 0.1	7.55 $\pm$ 0.0
Period	LRN* (Consumer)			BTC-USD		
	2025-11-03 ~2026-01-30			2025-06-05 ~2025-10-03		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	96.88	1.57	18.41	47.13	1.58	12.25
MACD	90.24	2.95	6.48	31.31	1.62	6.28
PEN	13.61 $\pm$ 8.0	0.47 $\pm$ 0.2	30.92 $\pm$ 5.0	-10.94 $\pm$ 8.3	-1.10 $\pm$ 0.1	29.69 $\pm$ 2.1
FININ	15.09 $\pm$ 7.0	0.51 $\pm$ 0.2	28.50 $\pm$ 1.1	4.71 $\pm$ 5.0	0.24 $\pm$ 0.2	19.08 $\pm$ 2.5
LLMFactor	22.34 $\pm$ 9.9	0.39 $\pm$ 0.2	26.26 $\pm$ 3.2	-0.14 $\pm$ 10.0	-0.02 $\pm$ 0.1	26.82 $\pm$ 6.0
FinMem	40.86 $\pm$ 13.8	1.02 $\pm$ 0.3	25.35 $\pm$ 1.0	45.80 $\pm$ 12.5	1.52 $\pm$ 0.3	19.26 $\pm$ 0.6
FinAgent	21.50 $\pm$ 10.7	0.62 $\pm$ 0.3	19.40 $\pm$ 1.9	42.19 $\pm$ 8.3	1.42 $\pm$ 0.2	21.79 $\pm$ 2.0
FinCon	118.10 $\pm$ 11.1	1.96 $\pm$ 0.2	11.91 $\pm$ 9.0	36.26 $\pm$ 9.0	1.29 $\pm$ 0.3	20.86 $\pm$ 4.8
TA	174.89 $\pm$ 28.3	2.94 $\pm$ 0.2	8.81 $\pm$ 4.6	56.07 $\pm$ 12.0	1.79 $\pm$ 0.2	7.63 $\pm$ 3.3
TiMi	158.31 $\pm$ 15.1	2.22 $\pm$ 0.1	6.95 $\pm$ 14.3	80.21 $\pm$ 14.3	2.37 $\pm$ 0.1	8.81 $\pm$ 3.2
FinRAQ	209.63 $\pm$ 3.7	3.07 $\pm$ 0.2	6.39 $\pm$ 0.2	121.3 $\pm$ 10.1	3.20 $\pm$ 0.1	5.56 $\pm$ 0.4

Table 18: Performance comparison under bull market conditions (B&H ARR > 20%; p < 0.05). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	NVDA			AAPL		
Period	2025-11-10 ~2026-02-06			2025-12-01 ~2026-02-07		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-25.40	-0.73	14.12	-24.56	-1.19	13.80
MACD	-31.58	-2.16	9.84	-10.42	-0.62	7.95
PEN	-73.68 $\pm$ 9.5	-3.59 $\pm$ 0.1	31.57 $\pm$ 2.5	-31.93 $\pm$ 6.3	-1.22 $\pm$ 0.1	18.53 $\pm$ 4.1
FININ	-61.02 $\pm$ 7.5	-2.46 $\pm$ 0.1	20.22 $\pm$ 5.8	-28.11 $\pm$ 5.8	-1.15 $\pm$ 0.1	23.17 $\pm$ 0.3
LLMFactor	-17.95 $\pm$ 12.7	-0.41 $\pm$ 0.1	13.74 $\pm$ 3.4	4.17 $\pm$ 8.1	0.21 $\pm$ 0.2	15.92 $\pm$ 3.8
FinMem	13.61 $\pm$ 7.2	0.47 $\pm$ 0.1	14.72 $\pm$ 4.9	32.66 $\pm$ 11.1	1.67 $\pm$ 0.1	13.84 $\pm$ 1.2
FinAgent	34.01 $\pm$ 9.9	1.17 $\pm$ 0.1	15.22 $\pm$ 4.8	39.27 $\pm$ 9.1	1.80 $\pm$ 0.1	14.93 $\pm$ 2.4
FinCon	-1.20 $\pm$ 5.1	0.09 $\pm$ 0.2	17.58 $\pm$ 9.9	34.56 $\pm$ 9.9	1.17 $\pm$ 0.1	15.55 $\pm$ 4.5
TA	53.86 $\pm$ 12.8	1.41 $\pm$ 0.1	9.92 $\pm$ 3.1	45.93 $\pm$ 10.6	2.00 $\pm$ 0.2	12.97 $\pm$ 3.4
TiMi	58.68 $\pm$ 11.5	1.53 $\pm$ 0.1	9.57 $\pm$ 3.3	60.93 $\pm$ 5.5	2.48 $\pm$ 0.1	11.35 $\pm$ 2.2
FinRAQ	76.55 $\pm$ 3.2	1.68 $\pm$ 0.1	9.16 $\pm$ 0.8	87.54 $\pm$ 1.0	2.76 $\pm$ 0.2	7.22 $\pm$ 0.3
	CIFR*			QBTS*		
Period	2025-12-01 ~2026-02-27			2025-12-01 ~2026-02-27		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-61.78	-0.44	35.96	-41.92	0.01	46.54
MACD	-60.94	-1.34	28.01	-50.62	-0.85	35.98
PEN	-73.10 $\pm$ 11.1	-0.75 $\pm$ 0.1	43.69 $\pm$ 1.8	-56.44 $\pm$ 12.0	-0.36 $\pm$ 0.1	50.40 $\pm$ 4.7
FININ	-42.10 $\pm$ 10.3	-0.57 $\pm$ 0.1	38.54 $\pm$ 3.0	-70.83 $\pm$ 10.2	-0.57 $\pm$ 0.1	59.10 $\pm$ 2.6
LLMFactor	33.56 $\pm$ 9.2	1.15 $\pm$ 0.2	31.34 $\pm$ 0.8	7.14 $\pm$ 10.1	0.56 $\pm$ 0.1	39.86 $\pm$ 1.4
FinMem	50.30 $\pm$ 5.0	1.16 $\pm$ 0.2	32.96 $\pm$ 1.7	6.62 $\pm$ 7.6	0.58 $\pm$ 0.2	52.24 $\pm$ 2.3
FinAgent	149.65 $\pm$ 8.8	1.22 $\pm$ 0.1	26.34 $\pm$ 1.6	58.86 $\pm$ 5.0	0.97 $\pm$ 0.1	43.34 $\pm$ 3.7
FinCon	136.34 $\pm$ 9.1	1.83 $\pm$ 0.1	27.79 $\pm$ 1.6	102.82 $\pm$ 7.5	1.02 $\pm$ 0.1	42.81 $\pm$ 0.4
TA	188.15 $\pm$ 9.6	2.52 $\pm$ 0.1	30.90 $\pm$ 1.0	114.80 $\pm$ 9.1	1.21 $\pm$ 0.1	39.46 $\pm$ 3.4
TiMi	161.72 $\pm$ 10.5	2.13 $\pm$ 0.2	26.67 $\pm$ 1.4	155.71 $\pm$ 10.3	1.04 $\pm$ 0.2	37.98 $\pm$ 0.8
FinRAQ	228.47 $\pm$ 4.5	2.77 $\pm$ 0.0	25.43 $\pm$ 0.5	202.63 $\pm$ 2.1	1.54 $\pm$ 0.0	29.60 $\pm$ 0.2

Table 19: Performance comparison of technology sector stocks under bear market conditions (B&H ARR < 20%; p < 0.05). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	JPM			MS		
Period	2025-02-13 ~2025-05-12			2025-02-18 ~2025-05-14		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-20.18	-0.45	24.42	-22.98	-0.32	29.05
MACD	-22.52	-0.86	18.32	42.88	1.28	16.01
PEN	-15.94 $\pm$ 12.0	-0.11 $\pm$ 0.1	23.99 $\pm$ 5.4	14.39 $\pm$ 5.9	0.48 $\pm$ 0.1	25.77 $\pm$ 12.0
FININ	-16.32 $\pm$ 10.5	-0.33 $\pm$ 0.0	14.73 $\pm$ 4.3	21.80 $\pm$ 4.3	0.73 $\pm$ 0.0	23.60 $\pm$ 4.7
LLMFactor	8.34 $\pm$ 5.4	0.80 $\pm$ 0.1	19.83 $\pm$ 4.1	26.22 $\pm$ 11.6	0.68 $\pm$ 0.1	27.42 $\pm$ 4.9
FinMem	30.39 $\pm$ 7.3	1.34 $\pm$ 0.1	15.56 $\pm$ 4.7	44.40 $\pm$ 6.9	1.02 $\pm$ 0.1	20.31 $\pm$ 4.2
FinAgent	32.27 $\pm$ 7.7	1.17 $\pm$ 0.2	14.12 $\pm$ 0.2	63.19 $\pm$ 12.7	1.19 $\pm$ 0.1	28.69 $\pm$ 6.8
FinCon	78.15 $\pm$ 7.0	1.22 $\pm$ 0.1	17.91 $\pm$ 2.3	68.50 $\pm$ 12.6	1.21 $\pm$ 0.1	28.12 $\pm$ 7.0
TA	90.47 $\pm$ 12.1	1.30 $\pm$ 0.1	13.71 $\pm$ 0.3	86.80 $\pm$ 14.1	1.85 $\pm$ 0.1	23.96 $\pm$ 1.1
TiMi	106.63 $\pm$ 16.4	2.14 $\pm$ 0.1	11.32 $\pm$ 4.9	110.41 $\pm$ 20.9	2.51 $\pm$ 0.1	19.35 $\pm$ 3.4
FinRAQ	154.63 $\pm$ 1.7	2.23 $\pm$ 0.2	9.39 $\pm$ 0.6	134.86 $\pm$ 5.2	2.91 $\pm$ 0.1	14.02 $\pm$ 1.0
	MARA*			WULF*		
Period	2025-12-01 ~2026-02-27			2025-11-07 ~2026-02-05		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-64.92	-0.74	46.03	-47.62	-0.21	29.56
MACD	16.48	0.53	14.18	-79.88	-1.78	32.17
PEN	63.58 $\pm$ 12.5	0.96 $\pm$ 0.1	23.89 $\pm$ 3.7	-48.30 $\pm$ 12.1	-0.22 $\pm$ 0.2	43.21 $\pm$ 6.4
FININ	-76.05 $\pm$ 12.0	-1.13 $\pm$ 0.3	53.24 $\pm$ 2.8	-79.07 $\pm$ 10.2	-1.15 $\pm$ 0.1	41.19 $\pm$ 1.0
LLMFactor	-29.53 $\pm$ 5.5	-1.03 $\pm$ 0.1	51.56 $\pm$ 5.2	-3.91 $\pm$ 10.3	-0.45 $\pm$ 0.2	37.97 $\pm$ 4.7
FinMem	35.52 $\pm$ 7.7	1.12 $\pm$ 0.3	22.35 $\pm$ 1.1	27.29 $\pm$ 7.7	0.84 $\pm$ 0.2	26.89 $\pm$ 1.6
FinAgent	33.78 $\pm$ 5.1	1.04 $\pm$ 0.2	27.36 $\pm$ 5.4	51.04 $\pm$ 8.3	1.47 $\pm$ 0.2	31.04 $\pm$ 5.1
FinCon	166.17 $\pm$ 8.1	1.87 $\pm$ 0.2	16.55 $\pm$ 1.0	98.86 $\pm$ 14.1	1.30 $\pm$ 0.2	30.84 $\pm$ 4.5
TA	150.06 $\pm$ 18.2	1.89 $\pm$ 0.1	25.13 $\pm$ 0.1	116.63 $\pm$ 13.9	1.26 $\pm$ 0.1	21.72 $\pm$ 4.2
TiMi	256.04 $\pm$ 14.7	2.69 $\pm$ 0.1	15.18 $\pm$ 1.5	149.33 $\pm$ 11.2	1.69 $\pm$ 0.1	18.42 $\pm$ 0.8
FinRAQ	303.97 $\pm$ 5.7	2.88 $\pm$ 0.2	13.86 $\pm$ 1.1	213.98 $\pm$ 3.8	2.78 $\pm$ 0.1	13.93 $\pm$ 1.4

Table 20: Performance comparison of financial sector stocks under bear market conditions (B&H ARR < 20%; p < 0.05). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	JNJ			UNH		
Period	2025-03-31 ~2025-06-26			2025-12-01 ~2026-02-27		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-27.71	-1.23	11.75	-31.20	-0.53	24.62
MACD	-25.04	-2.39	6.90	-52.75	-1.46	26.53
PEN	-41.76 $\pm$ 6.4	-2.09 $\pm$ 0.2	17.54 $\pm$ 1.6	-31.63 $\pm$ 11.9	-0.89 $\pm$ 0.1	29.95 $\pm$ 11.9
FININ	-19.39 $\pm$ 8.5	-0.74 $\pm$ 0.0	18.86 $\pm$ 3.1	-53.42 $\pm$ 9.4	-1.31 $\pm$ 0.1	30.43 $\pm$ 2.2
LLMFactor	-34.72 $\pm$ 13.4	-1.63 $\pm$ 0.1	17.60 $\pm$ 2.8	28.94 $\pm$ 11.5	0.70 $\pm$ 0.1	12.83 $\pm$ 3.1
FinMem	15.45 $\pm$ 9.5	0.61 $\pm$ 0.1	13.10 $\pm$ 1.9	36.55 $\pm$ 6.7	0.71 $\pm$ 0.1	19.61 $\pm$ 3.1
FinAgent	20.55 $\pm$ 7.5	1.15 $\pm$ 0.2	12.87 $\pm$ 0.3	29.70 $\pm$ 6.0	0.87 $\pm$ 0.2	13.67 $\pm$ 1.7
FinCon	22.87 $\pm$ 8.2	1.26 $\pm$ 0.1	13.83 $\pm$ 0.7	37.92 $\pm$ 8.2	0.86 $\pm$ 0.1	13.28 $\pm$ 1.2
TA	52.26 $\pm$ 11.0	1.65 $\pm$ 0.1	7.32 $\pm$ 4.7	44.98 $\pm$ 8.2	0.92 $\pm$ 0.1	17.53 $\pm$ 1.5
TiMi	67.76 $\pm$ 14.3	1.93 $\pm$ 0.1	6.95 $\pm$ 0.5	53.94 $\pm$ 9.6	1.15 $\pm$ 0.1	14.83 $\pm$ 3.3
FinRAQ	83.99 $\pm$ 4.0	2.47 $\pm$ 0.2	6.59 $\pm$ 0.2	72.92 $\pm$ 1.8	1.29 $\pm$ 0.0	12.58 $\pm$ 0.6
	IOVA*			IBRX*		
Period	2025-09-11 ~2025-12-05			2025-10-14 ~2026-01-09		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-32.55	-0.05	25.21	-26.47	-0.17	25.56
MACD	-62.98	-1.60	25.75	-39.15	-0.87	17.47
PEN	-53.85 $\pm$ 11.4	-1.96 $\pm$ 0.2	36.89 $\pm$ 4.7	-45.44 $\pm$ 17.4	-0.61 $\pm$ 0.2	23.74 $\pm$ 4.4
FININ	-92.45 $\pm$ 11.5	-2.41 $\pm$ 0.3	46.33 $\pm$ 1.1	-70.05 $\pm$ 14.6	-1.52 $\pm$ 0.4	34.64 $\pm$ 4.4
LLMFactor	-56.63 $\pm$ 7.5	-0.48 $\pm$ 0.3	40.11 $\pm$ 3.0	9.83 $\pm$ 5.5	0.24 $\pm$ 0.2	28.10 $\pm$ 3.2
FinMem	-15.69 $\pm$ 10.0	0.20 $\pm$ 0.2	33.29 $\pm$ 3.8	20.28 $\pm$ 11.9	1.15 $\pm$ 0.1	28.76 $\pm$ 1.4
FinAgent	88.57 $\pm$ 10.7	1.05 $\pm$ 0.2	28.03 $\pm$ 4.6	22.26 $\pm$ 11.0	1.47 $\pm$ 0.2	25.67 $\pm$ 2.1
FinCon	119.56 $\pm$ 5.7	1.27 $\pm$ 0.2	21.26 $\pm$ 0.0	-14.14 $\pm$ 10.3	0.07 $\pm$ 0.2	29.48 $\pm$ 5.3
TA	103.81 $\pm$ 15.2	1.20 $\pm$ 0.1	22.73 $\pm$ 1.6	54.87 $\pm$ 7.3	0.95 $\pm$ 0.1	20.26 $\pm$ 3.2
TiMi	144.03 $\pm$ 12.5	1.39 $\pm$ 0.2	18.76 $\pm$ 0.0	65.76 $\pm$ 11.1	1.06 $\pm$ 0.2	18.95 $\pm$ 1.6
FinRAQ	156.08 $\pm$ 1.7	1.64 $\pm$ 0.2	17.70 $\pm$ 0.6	74.19 $\pm$ 3.4	1.74 $\pm$ 0.0	15.57 $\pm$ 1.7

Table 21: Performance comparison of healthcare sector stocks under bear market conditions (B&H ARR < 20%; p < 0.05). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	XOM			CVX		
Period	2025-04-02 ~2025-06-30			2025-09-26 ~2025-12-22		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-30.16	-1.00	15.79	-20.58	-1.30	7.35
MACD	2.74	0.13	5.54	-21.09	-1.96	7.51
PEN	-37.03 $\pm$ 9.9	-1.32 $\pm$ 0.2	21.50 $\pm$ 3.4	21.88 $\pm$ 7.6	1.08 $\pm$ 0.1	9.15 $\pm$ 3.4
FININ	-32.15 $\pm$ 10.5	-1.47 $\pm$ 0.2	26.30 $\pm$ 6.3	7.21 $\pm$ 7.3	0.36 $\pm$ 0.3	15.60 $\pm$ 1.0
LLMFactor	-17.74 $\pm$ 10.1	-0.50 $\pm$ 0.2	13.26 $\pm$ 2.0	16.56 $\pm$ 2.0	0.83 $\pm$ 0.2	11.87 $\pm$ 4.5
FinMem	-23.76 $\pm$ 7.4	-0.73 $\pm$ 0.2	22.08 $\pm$ 3.3	43.14 $\pm$ 7.4	0.98 $\pm$ 0.1	7.68 $\pm$ 7.4
FinAgent	-8.35 $\pm$ 6.1	-0.17 $\pm$ 0.1	15.20 $\pm$ 2.3	35.89 $\pm$ 6.4	0.69 $\pm$ 0.2	8.03 $\pm$ 6.1
FinCon	90.94 $\pm$ 10.1	2.45 $\pm$ 0.2	14.72 $\pm$ 1.1	59.41 $\pm$ 1.1	1.01 $\pm$ 0.2	7.12 $\pm$ 2.0
TA	133.10 $\pm$ 9.5	2.73 $\pm$ 0.1	10.88 $\pm$ 2.8	58.02 $\pm$ 10.4	1.18 $\pm$ 0.2	7.41 $\pm$ 3.3
TiMi	138.64 $\pm$ 13.1	3.11 $\pm$ 0.2	7.61 $\pm$ 2.5	69.42 $\pm$ 11.7	1.36 $\pm$ 0.1	8.81 $\pm$ 0.7
FinRAQ	162.40 $\pm$ 3.0	3.25 $\pm$ 0.1	5.31 $\pm$ 0.4	82.59 $\pm$ 4.4	1.62 $\pm$ 0.1	5.20 $\pm$ 0.2
	RIG*			PTEN*		
Period	2025-04-02 ~2025-06-30			2025-07-24 ~2025-10-17		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-56.03	-0.55	32.59	-30.44	-0.61	17.29
MACD	122.78	1.78	16.27	-40.77	-1.62	15.63
PEN	-32.43 $\pm$ 10.8	-0.75 $\pm$ 0.2	33.92 $\pm$ 3.3	-73.55 $\pm$ 9.1	-2.74 $\pm$ 0.1	37.86 $\pm$ 9.1
FININ	-45.37 $\pm$ 14.0	-0.31 $\pm$ 0.1	38.99 $\pm$ 7.2	10.58 $\pm$ 6.6	0.70 $\pm$ 0.2	33.26 $\pm$ 1.0
LLMFactor	7.91 $\pm$ 8.5	0.28 $\pm$ 0.3	33.67 $\pm$ 4.4	0.73 $\pm$ 5.0	0.17 $\pm$ 0.3	35.96 $\pm$ 2.0
FinMem	48.60 $\pm$ 7.1	0.36 $\pm$ 0.4	41.94 $\pm$ 4.8	9.42 $\pm$ 9.7	0.28 $\pm$ 0.2	38.18 $\pm$ 0.9
FinAgent	14.98 $\pm$ 5.2	0.22 $\pm$ 0.2	39.80 $\pm$ 1.7	32.38 $\pm$ 9.1	0.63 $\pm$ 0.4	30.82 $\pm$ 2.7
FinCon	70.34 $\pm$ 16.7	1.53 $\pm$ 0.1	39.65 $\pm$ 5.9	68.12 $\pm$ 9.5	1.63 $\pm$ 0.1	15.30 $\pm$ 4.8
TA	83.05 $\pm$ 10.0	1.76 $\pm$ 0.2	24.35 $\pm$ 2.0	77.92 $\pm$ 8.9	1.89 $\pm$ 0.2	14.61 $\pm$ 4.5
TiMi	93.51 $\pm$ 10.8	1.80 $\pm$ 0.2	29.30 $\pm$ 3.7	107.02 $\pm$ 3.7	2.08 $\pm$ 0.1	14.40 $\pm$ 2.4
FinRAQ	126.43 $\pm$ 3.4	2.04 $\pm$ 0.2	13.92 $\pm$ 1.1	121.13 $\pm$ 9.9	2.23 $\pm$ 0.1	13.32 $\pm$ 1.2

Table 22: Performance comparison of energy sector stocks under bear market conditions (B&H ARR < 20%; p < 0.05). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	WMT			KO		
Period	2025-02-19 ~2025-05-15			2025-07-07 ~2025-09-30		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-25.50	-0.66	21.14	-22.18	-2.03	6.81
MACD	53.05	1.59	8.88	-1.90	-0.51	6.86
PEN	-35.17 $\pm$ 9.5	-1.45 $\pm$ 0.1	17.89 $\pm$ 3.7	-10.86 $\pm$ 5.5	-0.97 $\pm$ 0.2	5.23 $\pm$ 4.1
FININ	-51.56 $\pm$ 6.9	-1.82 $\pm$ 0.3	19.79 $\pm$ 3.5	1.81 $\pm$ 6.1	0.03 $\pm$ 0.4	5.52 $\pm$ 4.4
LLMFactor	25.45 $\pm$ 8.1	0.74 $\pm$ 0.2	17.91 $\pm$ 2.6	9.04 $\pm$ 5.1	0.25 $\pm$ 0.2	6.50 $\pm$ 1.4
FinMem	85.91 $\pm$ 9.0	1.81 $\pm$ 0.1	9.55 $\pm$ 4.6	15.26 $\pm$ 10.5	0.24 $\pm$ 0.2	7.52 $\pm$ 3.3
FinAgent	106.15 $\pm$ 13.9	2.14 $\pm$ 0.1	9.18 $\pm$ 0.8	17.63 $\pm$ 12.1	0.58 $\pm$ 0.2	6.67 $\pm$ 1.5
FinCon	122.77 $\pm$ 9.1	2.15 $\pm$ 0.1	9.14 $\pm$ 4.9	27.84 $\pm$ 9.2	1.50 $\pm$ 0.1	5.59 $\pm$ 4.1
TA	208.56 $\pm$ 8.4	2.39 $\pm$ 0.1	8.76 $\pm$ 0.3	35.49 $\pm$ 10.7	2.46 $\pm$ 0.1	5.08 $\pm$ 4.2
TiMi	228.75 $\pm$ 10.7	2.68 $\pm$ 0.2	7.78 $\pm$ 3.4	33.18 $\pm$ 3.4	2.25 $\pm$ 0.2	5.66 $\pm$ 1.5
FinRAQ	261.65 $\pm$ 4.7	2.80 $\pm$ 0.3	7.49 $\pm$ 0.2	45.85 $\pm$ 2.8	2.85 $\pm$ 0.2	4.45 $\pm$ 0.2
	LAUR*			LRN*		
Period	2024-05-28 ~2024-08-22			2025-10-28 ~2026-01-26		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-22.82	-0.72	12.41	-95.72	-1.82	60.26
MACD	-10.31	-0.38	12.01	53.26	2.00	7.31
PEN	6.58 $\pm$ 9.2	0.07 $\pm$ 0.1	16.39 $\pm$ 1.6	-99.09 $\pm$ 13.7	-3.20 $\pm$ 0.3	68.09 $\pm$ 5.0
FININ	5.30 $\pm$ 8.4	0.06 $\pm$ 0.2	16.89 $\pm$ 2.0	-95.67 $\pm$ 6.6	-1.81 $\pm$ 0.2	55.66 $\pm$ 1.3
LLMFactor	43.05 $\pm$ 11.3	1.22 $\pm$ 0.1	9.07 $\pm$ 2.8	10.75 $\pm$ 2.8	1.03 $\pm$ 0.2	9.09 $\pm$ 4.4
FinMem	57.16 $\pm$ 16.4	1.27 $\pm$ 0.1	8.98 $\pm$ 4.6	-2.63 $\pm$ 6.3	-0.35 $\pm$ 0.2	28.41 $\pm$ 5.9
FinAgent	69.66 $\pm$ 15.6	1.81 $\pm$ 0.2	9.33 $\pm$ 3.3	33.22 $\pm$ 9.5	1.80 $\pm$ 0.2	9.46 $\pm$ 3.3
FinCon	85.95 $\pm$ 8.4	1.93 $\pm$ 0.2	8.71 $\pm$ 1.1	34.14 $\pm$ 8.7	1.73 $\pm$ 1.1	7.25 $\pm$ 1.1
TA	90.50 $\pm$ 9.8	1.98 $\pm$ 0.2	8.43 $\pm$ 1.3	91.46 $\pm$ 13.7	2.22 $\pm$ 0.0	7.49 $\pm$ 4.1
TiMi	108.12 $\pm$ 9.6	1.92 $\pm$ 0.2	7.99 $\pm$ 3.4	114.11 $\pm$ 12.0	2.27 $\pm$ 0.2	8.84 $\pm$ 9.6
FinRAQ	133.65 $\pm$ 3.6	2.29 $\pm$ 0.1	5.78 $\pm$ 0.3	130.99 $\pm$ 2.2	2.38 $\pm$ 0.0	6.97 $\pm$ 0.0

Table 23: Performance comparison of consumer staples sector stocks under bear market conditions (B&H ARR < 20%; p < 0.05). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	BTC-USD		
Period	2025-12-30 ~2026-02-28		
	AR%↑	SR↑	MDD%↓
B&H	-68.23	-1.96	35.31
MACD	-16.87	-0.72	15.13
PEN	-86.16 $\pm$ 9.5	-3.61 $\pm$ 0.3	38.13 $\pm$ 3.7
FININ	-84.92 $\pm$ 16.7	-3.44 $\pm$ 0.1	38.31 $\pm$ 4.7
LLMFactor	2.23 $\pm$ 11.1	0.46 $\pm$ 0.1	27.93 $\pm$ 3.8
FinMem	4.28 $\pm$ 5.0	0.57 $\pm$ 0.3	25.75 $\pm$ 1.5
FinAgent	38.23 $\pm$ 8.3	1.02 $\pm$ 0.2	20.08 $\pm$ 3.7
FinCon	60.46 $\pm$ 12.2	1.10 $\pm$ 0.3	14.71 $\pm$ 3.4
TA	73.61 $\pm$ 15.2	1.28 $\pm$ 0.1	13.10 $\pm$ 3.5
TiMi	80.56 $\pm$ 11.7	1.37 $\pm$ 0.1	13.38 $\pm$ 2.4
FinRAQ	93.49 $\pm$ 4.1	1.48 $\pm$ 0.0	11.92 $\pm$ 0.3

Table 24: Performance comparison of BTC-USD under bear market conditions (B&H ARR < 20%; p < 0.05). Bold values indicate the best performance.

	NVDA			AAPL		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	-14.51	-0.6	36.72	-2.33	-0.19	33.35
MACD	-41.86	-1.72	40.96	3.07	0.12	16.22
FinMem	29.92	0.58	42.72	4.64	0.21	21.35
FinAgent	76.88	1.15	40.45	23.68	0.99	14.49
FinCon	70.20	1.06	27.90	26.32	1.54	14.18
TA	91.21	1.39	41.95	47.05	1.53	16.50
TiMi	102.87	1.48	39.06	43.21	1.48	15.92
FinRAQ	143.78	1.84	32.23	60.91	1.91	14.26
	META			TSLA		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
B&H	10.14	0.37	33.23	41.82	0.74	53.44
MACD	23.09	1.28	11.54	47.37	0.90	35.57
FinMem	14.91	0.47	21.95	52.65	0.79	44.65
FinAgent	28.63	1.11	19.24	138.09	1.17	47.64
FinCon	49.96	1.40	10.99	92.04	0.98	37.00
TA	37.83	1.01	24.23	124.07	1.09	42.10
TiMi	46.13	1.29	26.38	117.81	1.37	38.24
FinRAQ	63.12	2.02	14.28	191.63	1.67	36.07

Table 25: Performance comparison from July 1, 2024, to May 1, 2025, using GPT-4o with a knowledge cutoff predating the evaluation period to prevent information leakage. Bold values indicate the best performance.

	NVDA			AAPL		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	94.56 $\pm$ 1.7	1.56 $\pm$ 0.1	32.89 $\pm$ 3.1	62.17 $\pm$ 0.1	1.99 $\pm$ 0.2	21.92 $\pm$ 0.3
Self-consistency	95.88 $\pm$ 6.2	1.56 $\pm$ 0.1	32.39 $\pm$ 0.2	65.41 $\pm$ 0.8	1.85 $\pm$ 0.2	26.18 $\pm$ 0.0
Reflexion	96.57 $\pm$ 3.3	1.58 $\pm$ 0.0	46.89 $\pm$ 0.1	68.49 $\pm$ 4.8	1.91 $\pm$ 0.1	26.10 $\pm$ 1.3
MAD	98.53 $\pm$ 7.6	1.59 $\pm$ 0.1	53.47 $\pm$ 5.2	73.02 $\pm$ 2.4	2.01 $\pm$ 0.1	21.33 $\pm$ 1.7
UQ	100.07 $\pm$ 8.5	1.61 $\pm$ 0.1	45.18 $\pm$ 0.9	73.23 $\pm$ 4.7	2.01 $\pm$ 0.2	21.74 $\pm$ 0.5
FinRAQ	114.57 $\pm$ 4.3	1.78 $\pm$ 0.1	32.00 $\pm$ 1.1	76.20 $\pm$ 2.4	2.07 $\pm$ 0.0	17.87 $\pm$ 1.4
	CIFR*			QBTS*		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	214.25 $\pm$ 5.0	1.52 $\pm$ 0.1	64.96 $\pm$ 1.4	287.75 $\pm$ 21.4	1.64 $\pm$ 0.2	65.60 $\pm$ 6.6
Self-consistency	228.32 $\pm$ 2.2	1.54 $\pm$ 0.1	62.53 $\pm$ 0.6	289.51 $\pm$ 28.0	1.68 $\pm$ 0.0	75.24 $\pm$ 5.6
Reflexion	222.32 $\pm$ 17.6	1.55 $\pm$ 0.2	70.71 $\pm$ 4.1	288.27 $\pm$ 16.5	1.69 $\pm$ 0.1	79.65 $\pm$ 3.7
MAD	231.45 $\pm$ 21.6	1.55 $\pm$ 0.3	62.19 $\pm$ 5.6	301.21 $\pm$ 15.0	1.68 $\pm$ 0.1	82.50 $\pm$ 7.4
UQ	230.16 $\pm$ 15.9	1.51 $\pm$ 0.1	62.05 $\pm$ 2.1	296.13 $\pm$ 10.5	1.66 $\pm$ 0.0	79.28 $\pm$ 6.5
FinRAQ	238.36 $\pm$ 3.6	1.61 $\pm$ 0.0	61.21 $\pm$ 0.3	331.50 $\pm$ 5.3	1.70 $\pm$ 0.1	51.76 $\pm$ 3.7

Table 26: Performance comparison between Socratic Questioning (SQ) and other prompting strategies on technology sector stocks ($p < 0.05$). ‘*’ denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	JPM			MS		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	87.28 $\pm$ 1.8	2.56 $\pm$ 0.2	15.41 $\pm$ 0.9	108.78 $\pm$ 7.3	2.01 $\pm$ 0.1	16.65 $\pm$ 1.2
Self-consistency	85.92 $\pm$ 5.0	2.53 $\pm$ 0.1	18.54 $\pm$ 0.5	106.20 $\pm$ 8.6	2.54 $\pm$ 0.2	16.30 $\pm$ 1.1
Reflexion	89.60 $\pm$ 2.2	2.61 $\pm$ 0.2	16.37 $\pm$ 0.4	108.78 $\pm$ 3.3	2.58 $\pm$ 0.2	16.61 $\pm$ 0.5
MAD	91.16 $\pm$ 4.9	2.65 $\pm$ 0.0	13.71 $\pm$ 0.1	109.80 $\pm$ 8.0	2.60 $\pm$ 0.1	23.12 $\pm$ 1.3
UQ	95.67 $\pm$ 3.8	2.74 $\pm$ 0.1	21.74 $\pm$ 2.0	112.78 $\pm$ 4.4	2.64 $\pm$ 0.2	16.35 $\pm$ 1.0
FinRAQ	104.67 $\pm$ 4.2	2.93 $\pm$ 0.2	12.08 $\pm$ 0.8	113.29 $\pm$ 5.0	2.66 $\pm$ 0.2	15.56 $\pm$ 1.1
	MARA*			WULF*		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	44.79 $\pm$ 4.3	0.85 $\pm$ 0.0	73.76 $\pm$ 5.6	221.29 $\pm$ 3.7	1.56 $\pm$ 0.1	60.85 $\pm$ 0.2
Self-consistency	43.32 $\pm$ 0.9	0.82 $\pm$ 0.0	64.90 $\pm$ 4.6	219.47 $\pm$ 8.0	1.54 $\pm$ 0.1	55.16 $\pm$ 2.2
Reflexion	40.61 $\pm$ 1.8	0.80 $\pm$ 0.0	71.47 $\pm$ 5.2	216.30 $\pm$ 7.5	1.53 $\pm$ 0.0	58.79 $\pm$ 4.2
MAD	47.20 $\pm$ 1.0	0.82 $\pm$ 0.0	67.94 $\pm$ 2.3	231.83 $\pm$ 5.1	1.54 $\pm$ 0.0	53.79 $\pm$ 4.0
UQ	46.75 $\pm$ 0.1	0.80 $\pm$ 0.0	70.26 $\pm$ 2.0	234.53 $\pm$ 5.0	1.57 $\pm$ 0.1	66.19 $\pm$ 2.8
FinRAQ	48.04 $\pm$ 2.0	0.87 $\pm$ 0.0	55.72 $\pm$ 4.5	245.88 $\pm$ 1.2	1.68 $\pm$ 0.1	51.46 $\pm$ 4.5

Table 27: Performance comparison between Socratic Questioning (SQ) and other prompting strategies on financial sector stocks ($p < 0.05$). ‘**’ denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	JNJ			UNH		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	68.26 $\pm$ 0.3	2.93 $\pm$ 0.2	18.58 $\pm$ 0.3	55.56 $\pm$ 0.4	1.23 $\pm$ 0.0	24.68 $\pm$ 2.0
Self-consistency	74.69 $\pm$ 4.9	1.96 $\pm$ 0.1	17.61 $\pm$ 0.7	64.77 $\pm$ 0.4	1.38 $\pm$ 0.1	36.75 $\pm$ 2.0
Reflexion	69.30 $\pm$ 0.4	1.86 $\pm$ 0.0	20.43 $\pm$ 1.9	68.96 $\pm$ 6.3	1.43 $\pm$ 0.0	45.55 $\pm$ 4.1
MAD	76.90 $\pm$ 3.5	2.01 $\pm$ 0.0	17.47 $\pm$ 1.1	65.20 $\pm$ 3.6	1.37 $\pm$ 0.1	27.64 $\pm$ 2.2
UQ	76.93 $\pm$ 1.6	2.01 $\pm$ 0.2	32.28 $\pm$ 2.7	67.48 $\pm$ 6.6	1.45 $\pm$ 0.0	32.17 $\pm$ 0.6
FinRAQ	79.47 $\pm$ 2.3	3.04 $\pm$ 0.3	11.60 $\pm$ 0.5	73.81 $\pm$ 1.5	1.50 $\pm$ 0.0	22.65 $\pm$ 1.6
	IOVA*			IBRX*		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	65.95 $\pm$ 1.8	0.98 $\pm$ 0.1	57.26 $\pm$ 4.7	109.07 $\pm$ 0.4	1.27 $\pm$ 0.1	71.50 $\pm$ 3.9
Self-consistency	60.86 $\pm$ 4.2	0.93 $\pm$ 0.0	67.61 $\pm$ 4.0	110.68 $\pm$ 4.9	1.22 $\pm$ 0.1	61.91 $\pm$ 3.4
Reflexion	65.70 $\pm$ 3.2	1.01 $\pm$ 0.1	66.10 $\pm$ 1.6	110.44 $\pm$ 10.5	1.25 $\pm$ 0.1	77.92 $\pm$ 7.1
MAD	66.35 $\pm$ 0.0	1.02 $\pm$ 0.0	57.08 $\pm$ 3.2	113.32 $\pm$ 5.9	1.26 $\pm$ 0.0	82.11 $\pm$ 4.8
UQ	69.07 $\pm$ 4.9	1.01 $\pm$ 0.0	74.94 $\pm$ 0.7	112.64 $\pm$ 1.7	1.21 $\pm$ 0.1	78.55 $\pm$ 0.3
FinRAQ	74.76 $\pm$ 2.3	1.05 $\pm$ 0.2	55.27 $\pm$ 5.2	121.40 $\pm$ 3.7	1.31 $\pm$ 0.2	58.14 $\pm$ 3.6

Table 28: Performance comparison between Socratic Questioning (SQ) and other prompting strategies on healthcare sector stocks ($p < 0.05$). ‘**’ denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	XOM			CVX		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	71.30 $\pm$ 0.3	2.22 $\pm$ 0.2	16.35 $\pm$ 0.3	54.66 $\pm$ 2.2	2.08 $\pm$ 0.1	18.84 $\pm$ 1.8
Self-consistency	75.18 $\pm$ 4.6	2.56 $\pm$ 0.0	13.76 $\pm$ 0.6	60.62 $\pm$ 4.0	2.16 $\pm$ 0.1	20.27 $\pm$ 0.1
Reflexion	73.19 $\pm$ 3.7	2.50 $\pm$ 0.2	14.48 $\pm$ 0.4	60.82 $\pm$ 5.3	2.17 $\pm$ 0.2	16.85 $\pm$ 0.0
MAD	79.16 $\pm$ 7.2	2.66 $\pm$ 0.2	12.80 $\pm$ 0.6	57.69 $\pm$ 5.1	2.08 $\pm$ 0.2	15.65 $\pm$ 0.0
UQ	79.85 $\pm$ 7.7	2.68 $\pm$ 0.2	14.23 $\pm$ 0.1	64.05 $\pm$ 1.8	2.26 $\pm$ 0.2	15.95 $\pm$ 1.2
FinRAQ	86.98 $\pm$ 5.3	2.86 $\pm$ 0.0	12.08 $\pm$ 0.1	67.81 $\pm$ 3.4	2.31 $\pm$ 0.2	15.33 $\pm$ 0.7
	RIG*			PTEN*		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	79.85 $\pm$ 3.6	1.28 $\pm$ 0.0	42.53 $\pm$ 2.4	56.45 $\pm$ 5.3	1.10 $\pm$ 0.1	40.71 $\pm$ 1.9
Self-consistency	78.17 $\pm$ 2.7	1.27 $\pm$ 0.1	44.38 $\pm$ 3.6	58.07 $\pm$ 2.1	1.12 $\pm$ 0.1	54.11 $\pm$ 5.1
Reflexion	78.84 $\pm$ 0.2	1.28 $\pm$ 0.0	46.43 $\pm$ 0.3	57.43 $\pm$ 5.3	1.11 $\pm$ 0.0	41.46 $\pm$ 3.8
MAD	81.12 $\pm$ 1.7	1.29 $\pm$ 0.0	43.41 $\pm$ 0.3	62.99 $\pm$ 5.4	1.17 $\pm$ 0.1	57.54 $\pm$ 1.2
UQ	82.38 $\pm$ 2.3	1.29 $\pm$ 0.1	33.40 $\pm$ 3.2	59.47 $\pm$ 3.8	1.13 $\pm$ 0.0	47.03 $\pm$ 3.4
FinRAQ	87.36 $\pm$ 3.9	1.32 $\pm$ 0.0	30.57 $\pm$ 2.7	65.44 $\pm$ 6.0	1.21 $\pm$ 0.1	25.38 $\pm$ 1.2

Table 29: Performance comparison between Socratic Questioning (SQ) and other prompting strategies on energy sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	WMT			KO		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	114.23 $\pm$ 6.4	3.44 $\pm$ 0.2	17.95 $\pm$ 0.9	69.81 $\pm$ 1.9	3.39 $\pm$ 0.0	7.06 $\pm$ 0.5
Self-consistency	120.06 $\pm$ 5.0	3.57 $\pm$ 0.0	10.73 $\pm$ 0.0	71.60 $\pm$ 6.0	3.46 $\pm$ 0.2	8.44 $\pm$ 0.1
Reflexion	120.17 $\pm$ 1.6	3.57 $\pm$ 0.1	11.63 $\pm$ 0.9	68.40 $\pm$ 2.6	3.33 $\pm$ 0.2	11.53 $\pm$ 0.3
MAD	128.80 $\pm$ 3.6	3.75 $\pm$ 0.2	14.44 $\pm$ 0.8	74.12 $\pm$ 4.1	3.56 $\pm$ 0.1	8.52 $\pm$ 0.2
UQ	122.52 $\pm$ 9.4	3.62 $\pm$ 0.3	12.78 $\pm$ 0.9	73.68 $\pm$ 1.5	3.54 $\pm$ 0.2	7.98 $\pm$ 0.5
FinRAQ	137.57 $\pm$ 5.8	3.94 $\pm$ 0.1	10.57 $\pm$ 0.5	77.82 $\pm$ 3.6	3.64 $\pm$ 0.2	6.96 $\pm$ 0.3
	LAUR*			LRN*		
	AR%↑	SR↑	MDD%↓	AR%↑	SR↑	MDD%↓
CoT	201.17 $\pm$ 3.0	3.87 $\pm$ 0.3	16.00 $\pm$ 0.7	74.76 $\pm$ 2.6	1.31 $\pm$ 0.1	60.89 $\pm$ 5.8
Self-consistency	199.34 $\pm$ 0.9	3.86 $\pm$ 0.3	10.94 $\pm$ 0.9	70.92 $\pm$ 5.1	1.04 $\pm$ 0.0	54.95 $\pm$ 1.4
Reflexion	201.95 $\pm$ 0.6	3.88 $\pm$ 0.2	22.95 $\pm$ 0.2	73.67 $\pm$ 2.7	1.30 $\pm$ 0.1	55.87 $\pm$ 2.1
MAD	207.51 $\pm$ 6.8	3.95 $\pm$ 0.2	12.87 $\pm$ 0.6	79.68 $\pm$ 1.4	1.35 $\pm$ 0.0	55.66 $\pm$ 3.2
UQ	209.17 $\pm$ 10.2	3.97 $\pm$ 0.3	13.64 $\pm$ 0.0	81.71 $\pm$ 5.2	1.28 $\pm$ 0.0	54.64 $\pm$ 0.7
FinRAQ	218.19 $\pm$ 5.1	4.07 $\pm$ 0.2	10.46 $\pm$ 0.4	84.68 $\pm$ 3.8	1.38 $\pm$ 0.1	39.75 $\pm$ 1.1

Table 30: Performance comparison between Socratic Questioning (SQ) and other prompting strategies on consumer staples sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	BTC-USD		
	AR%↑	SR↑	MDD%↓
CoT	108.82 $\pm$ 6.2	1.71 $\pm$ 0.2	42.34 $\pm$ 3.2
Self-consistency	110.84 $\pm$ 1.8	1.73 $\pm$ 0.0	35.13 $\pm$ 2.4
Reflexion	113.72 $\pm$ 1.0	1.76 $\pm$ 0.0	42.37 $\pm$ 4.2
MAD	114.02 $\pm$ 2.2	1.76 $\pm$ 0.1	28.61 $\pm$ 2.5
UQ	114.26 $\pm$ 1.0	1.76 $\pm$ 0.1	20.53 $\pm$ 0.6
FinRAQ	116.38 $\pm$ 2.6	1.78 $\pm$ 0.1	18.08 $\pm$ 0.3

Table 31: Performance comparison between Socratic Questioning (SQ) and other prompting strategies baselines on BTC-USD ($p < 0.05$). Bold values indicate the best performance.

	NVDA			AAPL		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	45.77 $\pm$ 4.1	0.97 $\pm$ 0.0	44.17 $\pm$ 1.4	44.79 $\pm$ 0.5	1.38 $\pm$ 0.1	35.95 $\pm$ 0.8
w/o $\mathcal{L}_R$	57.96 $\pm$ 3.4	1.13 $\pm$ 0.0	40.17 $\pm$ 2.9	47.05 $\pm$ 2.1	1.43 $\pm$ 0.0	32.57 $\pm$ 1.5
w/o $\mathcal{L}_C$	112.75 $\pm$ 4.2	1.73 $\pm$ 0.2	35.46 $\pm$ 1.9	64.23 $\pm$ 1.0	1.82 $\pm$ 0.1	24.71 $\pm$ 0.7
w/o Finetuning	114.57 $\pm$ 4.3	1.78 $\pm$ 0.1	32.00 $\pm$ 1.1	76.20 $\pm$ 2.4	2.07 $\pm$ 0.0	17.87 $\pm$ 1.4
FinRAQ	129.53 $\pm$ 4.5	1.89 $\pm$ 0.1	31.84 $\pm$ 3.4	86.19 $\pm$ 3.6	2.27 $\pm$ 0.1	9.35 $\pm$ 3.8

	CIFR*			QBTS*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	104.69 $\pm$ 9.5	1.20 $\pm$ 0.1	78.97 $\pm$ 0.5	112.42 $\pm$ 6.9	1.25 $\pm$ 0.1	82.32 $\pm$ 1.8
w/o $\mathcal{L}_R$	130.25 $\pm$ 4.3	1.28 $\pm$ 0.2	75.43 $\pm$ 6.6	139.44 $\pm$ 4.0	1.36 $\pm$ 0.2	72.50 $\pm$ 3.8
w/o $\mathcal{L}_C$	228.88 $\pm$ 12.1	1.61 $\pm$ 0.3	63.69 $\pm$ 5.0	327.58 $\pm$ 17.2	1.73 $\pm$ 0.2	56.14 $\pm$ 6.6
w/o Finetuning	238.36 $\pm$ 3.6	1.61 $\pm$ 0.0	61.21 $\pm$ 0.3	331.50 $\pm$ 5.3	1.70 $\pm$ 0.1	51.76 $\pm$ 3.7
FinRAQ	249.69 $\pm$ 2.7	1.67 $\pm$ 0.1	57.14 $\pm$ 5.0	346.06 $\pm$ 3.6	1.81 $\pm$ 0.0	50.66 $\pm$ 3.8

Table 32: Component-wise ablation results on technology sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	JPM			MS		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	66.05 $\pm$ 1.4	2.07 $\pm$ 0.1	18.27 $\pm$ 0.3	69.09 $\pm$ 1.3	1.85 $\pm$ 0.0	20.04 $\pm$ 1.2
w/o $\mathcal{L}_R$	79.21 $\pm$ 2.8	2.38 $\pm$ 0.2	16.71 $\pm$ 0.8	74.57 $\pm$ 2.6	1.96 $\pm$ 0.1	19.47 $\pm$ 0.4
w/o $\mathcal{L}_C$	94.33 $\pm$ 9.2	2.42 $\pm$ 0.1	15.45 $\pm$ 0.9	106.63 $\pm$ 3.1	2.57 $\pm$ 0.0	16.00 $\pm$ 0.3
w/o Finetuning	104.67 $\pm$ 4.2	2.93 $\pm$ 0.2	12.08 $\pm$ 0.8	113.29 $\pm$ 5.0	2.66 $\pm$ 0.2	15.56 $\pm$ 1.1
FinRAQ	114.95 $\pm$ 2.2	3.13 $\pm$ 0.1	9.43 $\pm$ 4.1	126.98 $\pm$ 3.1	2.87 $\pm$ 0.0	15.02 $\pm$ 1.3

	MARA*			WULF*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	8.56 $\pm$ 0.8	0.53 $\pm$ 0.0	83.06 $\pm$ 1.2	59.22 $\pm$ 3.2	0.96 $\pm$ 0.0	78.21 $\pm$ 7.8
w/o $\mathcal{L}_R$	11.46 $\pm$ 0.9	0.55 $\pm$ 0.1	71.21 $\pm$ 3.4	75.84 $\pm$ 0.6	1.10 $\pm$ 0.1	70.83 $\pm$ 4.6
w/o $\mathcal{L}_C$	45.28 $\pm$ 3.3	0.85 $\pm$ 0.1	60.28 $\pm$ 4.9	235.54 $\pm$ 23.0	1.69 $\pm$ 0.1	64.90 $\pm$ 2.5
w/o Finetuning	48.04 $\pm$ 2.0	0.87 $\pm$ 0.0	55.72 $\pm$ 4.5	245.88 $\pm$ 1.2	1.68 $\pm$ 0.1	51.46 $\pm$ 4.5
FinRAQ	56.68 $\pm$ 4.8	0.93 $\pm$ 0.1	42.90 $\pm$ 1.6	255.15 $\pm$ 3.8	1.65 $\pm$ 0.1	50.17 $\pm$ 4.3

Table 33: Component-wise ablation results on financial sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	JNJ			UNH		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	57.76 $\pm$ 0.5	2.56 $\pm$ 0.2	11.16 $\pm$ 0.8	6.53 $\pm$ 0.6	0.32 $\pm$ 0.0	48.96 $\pm$ 1.6
w/o $\mathcal{L}_R$	58.16 $\pm$ 5.1	2.57 $\pm$ 0.1	12.82 $\pm$ 0.7	34.97 $\pm$ 0.4	0.89 $\pm$ 0.0	39.91 $\pm$ 3.6
w/o $\mathcal{L}_C$	65.37 $\pm$ 2.4	2.78 $\pm$ 0.1	12.18 $\pm$ 0.7	69.53 $\pm$ 2.4	1.28 $\pm$ 0.1	20.56 $\pm$ 0.3
w/o Finetuning	79.47 $\pm$ 2.3	3.04 $\pm$ 0.3	11.60 $\pm$ 0.5	73.81 $\pm$ 1.5	1.50 $\pm$ 0.0	22.65 $\pm$ 1.6
FinRAQ	89.51 $\pm$ 4.2	3.64 $\pm$ 0.0	9.12 $\pm$ 0.2	78.89 $\pm$ 6.2	1.57 $\pm$ 0.0	20.84 $\pm$ 0.4

	IOVA*			IBRX*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	20.29 $\pm$ 1.5	0.64 $\pm$ 0.0	71.28 $\pm$ 2.8	27.95 $\pm$ 0.0	0.78 $\pm$ 0.0	68.76 $\pm$ 4.5
w/o $\mathcal{L}_R$	33.66 $\pm$ 3.0	0.78 $\pm$ 0.0	68.07 $\pm$ 3.7	37.76 $\pm$ 3.0	0.80 $\pm$ 0.1	66.18 $\pm$ 3.1
w/o $\mathcal{L}_C$	68.58 $\pm$ 4.5	1.01 $\pm$ 0.1	55.41 $\pm$ 5.9	110.41 $\pm$ 2.3	1.24 $\pm$ 0.0	69.38 $\pm$ 2.9
w/o Finetuning	74.76 $\pm$ 2.3	1.05 $\pm$ 0.2	55.27 $\pm$ 5.2	121.40 $\pm$ 3.7	1.31 $\pm$ 0.2	58.14 $\pm$ 3.6
FinRAQ	77.21 $\pm$ 1.5	1.10 $\pm$ 0.0	50.77 $\pm$ 3.6	124.27 $\pm$ 1.5	1.32 $\pm$ 0.0	53.76 $\pm$ 4.0

Table 34: Component-wise ablation results on healthcare sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	XOM			CVX		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	59.28 $\pm$ 4.4	2.12 $\pm$ 0.2	15.85 $\pm$ 0.5	30.29 $\pm$ 2.0	1.21 $\pm$ 0.1	22.97 $\pm$ 1.1
w/o $\mathcal{L}_R$	63.57 $\pm$ 6.2	2.24 $\pm$ 0.1	15.14 $\pm$ 0.2	32.99 $\pm$ 0.9	1.30 $\pm$ 0.1	18.15 $\pm$ 0.6
w/o $\mathcal{L}_C$	75.16 $\pm$ 2.9	2.56 $\pm$ 0.1	12.58 $\pm$ 0.5	55.60 $\pm$ 3.1	1.92 $\pm$ 0.1	17.09 $\pm$ 0.1
w/o Finetuning	86.98 $\pm$ 5.3	2.86 $\pm$ 0.0	12.08 $\pm$ 0.1	67.81 $\pm$ 3.4	2.31 $\pm$ 0.2	15.33 $\pm$ 0.7
FinRAQ	94.78 $\pm$ 3.6	3.05 $\pm$ 0.0	9.24 $\pm$ 0.5	71.68 $\pm$ 0.6	2.47 $\pm$ 0.0	14.98 $\pm$ 1.9
	RIG*			PTEN*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	0.47 $\pm$ 3.0	0.26 $\pm$ 0.0	66.83 $\pm$ 4.6	11.07 $\pm$ 4.0	0.42 $\pm$ 0.0	59.28 $\pm$ 2.7
w/o $\mathcal{L}_R$	40.72 $\pm$ 1.6	0.85 $\pm$ 0.0	60.42 $\pm$ 3.5	23.33 $\pm$ 3.6	0.63 $\pm$ 0.1	54.78 $\pm$ 0.8
w/o $\mathcal{L}_C$	77.93 $\pm$ 7.6	1.26 $\pm$ 0.1	57.50 $\pm$ 5.7	64.53 $\pm$ 4.2	1.20 $\pm$ 0.1	46.26 $\pm$ 4.3
w/o Finetuning	87.36 $\pm$ 3.9	1.32 $\pm$ 0.0	30.57 $\pm$ 2.7	65.44 $\pm$ 6.0	1.21 $\pm$ 0.1	25.38 $\pm$ 1.2
FinRAQ	97.15 $\pm$ 4.5	1.44 $\pm$ 0.1	28.70 $\pm$ 0.2	71.37 $\pm$ 2.1	1.29 $\pm$ 0.1	34.45 $\pm$ 2.4

Table 35: Component-wise ablation results on energy sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	WMT			KO		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	55.93 $\pm$ 4.9	1.99 $\pm$ 0.1	23.46 $\pm$ 0.6	36.19 $\pm$ 0.3	1.92 $\pm$ 0.0	11.95 $\pm$ 0.9
w/o $\mathcal{L}_R$	65.93 $\pm$ 0.1	2.27 $\pm$ 0.1	16.17 $\pm$ 1.2	48.19 $\pm$ 1.6	2.48 $\pm$ 0.1	10.69 $\pm$ 0.8
w/o $\mathcal{L}_C$	123.13 $\pm$ 2.5	3.64 $\pm$ 0.1	13.96 $\pm$ 0.5	72.07 $\pm$ 5.0	3.47 $\pm$ 0.0	8.05 $\pm$ 0.7
w/o Finetuning	137.57 $\pm$ 5.8	3.94 $\pm$ 0.1	10.57 $\pm$ 0.5	77.82 $\pm$ 3.6	3.64 $\pm$ 0.2	6.96 $\pm$ 0.3
FinRAQ	140.92 $\pm$ 2.6	4.00 $\pm$ 0.2	8.64 $\pm$ 0.2	81.95 $\pm$ 4.8	3.86 $\pm$ 0.1	6.86 $\pm$ 1.3
	LAUR*			LRN*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	66.02 $\pm$ 4.4	1.78 $\pm$ 0.1	17.04 $\pm$ 1.0	18.83 $\pm$ 4.2	0.51 $\pm$ 0.0	54.67 $\pm$ 1.0
w/o $\mathcal{L}_R$	70.04 $\pm$ 4.7	1.86 $\pm$ 0.1	16.11 $\pm$ 0.8	32.05 $\pm$ 1.9	0.80 $\pm$ 0.1	36.32 $\pm$ 0.9
w/o $\mathcal{L}_C$	205.40 $\pm$ 5.2	3.92 $\pm$ 0.3	10.75 $\pm$ 0.7	66.10 $\pm$ 6.3	1.09 $\pm$ 0.1	46.34 $\pm$ 4.5
w/o Finetuning	218.19 $\pm$ 5.1	4.07 $\pm$ 0.2	10.46 $\pm$ 0.4	84.68 $\pm$ 3.8	1.38 $\pm$ 0.1	39.75 $\pm$ 1.1
FinRAQ	229.92 $\pm$ 3.4	4.20 $\pm$ 0.2	9.13 $\pm$ 0.2	87.39 $\pm$ 5.1	1.39 $\pm$ 0.1	30.72 $\pm$ 2.9

Table 36: Component-wise ablation results on consumer staples sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	BTC-USD		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
w/o NewsRank	27.54 $\pm$ 2.1	0.70 $\pm$ 0.0	41.34 $\pm$ 3.0
w/o $\mathcal{L}_R$	47.35 $\pm$ 1.0	1.00 $\pm$ 0.0	35.29 $\pm$ 1.7
w/o $\mathcal{L}_C$	114.40 $\pm$ 7.0	1.76 $\pm$ 0.2	20.91 $\pm$ 1.1
w/o Finetuning	116.38 $\pm$ 2.6	1.78 $\pm$ 0.1	18.08 $\pm$ 0.3
FinRAQ	128.36 $\pm$ 1.4	1.89 $\pm$ 0.1	15.97 $\pm$ 0.6

Table 37: Component-wise ablation results on BTC-USD ($p < 0.05$). Best results are highlighted in bold.

	NVDA			AAPL		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	28.26 $\pm$ 2.1	0.71 $\pm$ 0.1	34.75 $\pm$ 1.7	18.10 $\pm$ 1.0	0.65 $\pm$ 0.0	29.27 $\pm$ 1.6
Discrete	122.45 $\pm$ 1.2	1.83 $\pm$ 0.2	32.09 $\pm$ 0.9	78.34 $\pm$ 7.6	2.12 $\pm$ 0.1	15.30 $\pm$ 0.2
Continuous (Ours)	129.53 $\pm$ 4.5	1.89 $\pm$ 0.1	31.84 $\pm$ 3.4	86.19 $\pm$ 3.6	2.27 $\pm$ 0.1	9.35 $\pm$ 3.8
	CIFR*			QBTS*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	38.00 $\pm$ 0.6	0.84 $\pm$ 0.1	85.27 $\pm$ 0.0	38.50 $\pm$ 2.0	0.92 $\pm$ 0.0	91.62 $\pm$ 5.9
Discrete	246.50 $\pm$ 12.6	1.65 $\pm$ 0.0	62.19 $\pm$ 4.6	303.09 $\pm$ 10.5	1.74 $\pm$ 0.1	55.41 $\pm$ 4.9
Continuous (Ours)	249.69 $\pm$ 2.7	1.67 $\pm$ 0.1	57.14 $\pm$ 5.0	346.06 $\pm$ 3.6	1.81 $\pm$ 0.0	50.66 $\pm$ 3.8

Table 38: Performance comparison of scoring strategies on technology sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	JPM			MS		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	21.98 $\pm$ 0.8	0.84 $\pm$ 0.0	36.64 $\pm$ 1.3	23.95 $\pm$ 0.3	0.80 $\pm$ 0.0	31.50 $\pm$ 1.4
Discrete	105.83 $\pm$ 4.2	2.96 $\pm$ 0.1	15.24 $\pm$ 0.4	123.83 $\pm$ 6.7	2.82 $\pm$ 0.2	15.91 $\pm$ 1.1
Continuous (Ours)	114.95 $\pm$ 2.2	3.13 $\pm$ 0.1	9.43 $\pm$ 4.1	126.98 $\pm$ 3.1	2.87 $\pm$ 0.0	15.02 $\pm$ 1.3
	MARA*			WULF*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	-18.24 $\pm$ 1.6	0.21 $\pm$ 0.0	89.10 $\pm$ 7.6	8.79 $\pm$ 0.4	0.64 $\pm$ 0.0	68.06 $\pm$ 6.3
Discrete	52.67 $\pm$ 4.1	0.91 $\pm$ 0.1	51.20 $\pm$ 3.1	254.77 $\pm$ 4.3	1.63 $\pm$ 0.1	56.96 $\pm$ 2.4
Continuous (Ours)	56.68 $\pm$ 4.8	0.93 $\pm$ 0.1	42.90 $\pm$ 1.6	255.15 $\pm$ 3.8	1.65 $\pm$ 0.1	50.17 $\pm$ 4.3

Table 39: Performance comparison of scoring strategies on financial sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	JNJ			UNH		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	9.92 $\pm$ 0.9	0.51 $\pm$ 0.0	35.68 $\pm$ 1.1	9.89 $\pm$ 0.4	0.39 $\pm$ 0.0	48.66 $\pm$ 1.7
Discrete	83.32 $\pm$ 3.7	3.44 $\pm$ 0.2	10.40 $\pm$ 0.6	73.30 $\pm$ 3.2	1.49 $\pm$ 0.1	23.28 $\pm$ 2.1
Continuous (Ours)	89.51 $\pm$ 4.2	3.64 $\pm$ 0.0	9.12 $\pm$ 0.2	78.89 $\pm$ 6.2	1.57 $\pm$ 0.0	20.84 $\pm$ 0.4
	IOVA*			IBRX*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	13.34 $\pm$ 0.9	0.56 $\pm$ 0.0	68.70 $\pm$ 3.1	39.05 $\pm$ 1.7	0.82 $\pm$ 0.0	81.41 $\pm$ 2.0
Discrete	73.19 $\pm$ 1.5	1.04 $\pm$ 0.0	56.92 $\pm$ 0.8	118.01 $\pm$ 8.3	1.29 $\pm$ 0.1	60.67 $\pm$ 5.8
Continuous (Ours)	77.21 $\pm$ 1.5	1.10 $\pm$ 0.0	50.77 $\pm$ 3.6	124.27 $\pm$ 1.5	1.32 $\pm$ 0.0	53.76 $\pm$ 4.0

Table 40: Performance comparison of scoring strategies on healthcare sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	XOM			CVX		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	18.67 $\pm$ 1.3	0.79 $\pm$ 0.0	22.07 $\pm$ 1.4	13.85 $\pm$ 0.1	0.60 $\pm$ 0.0	28.86 $\pm$ 0.2
Discrete	87.96 $\pm$ 6.4	2.89 $\pm$ 0.0	11.32 $\pm$ 0.3	67.39 $\pm$ 5.1	2.35 $\pm$ 0.2	17.21 $\pm$ 1.0
Continuous (Ours)	94.78 $\pm$ 3.6	3.05 $\pm$ 0.0	9.24 $\pm$ 0.5	71.68 $\pm$ 0.6	2.47 $\pm$ 0.0	14.98 $\pm$ 1.9
	RIG*			PTEN*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	28.51 $\pm$ 2.4	0.69 $\pm$ 0.1	57.77 $\pm$ 4.5	0.44 $\pm$ 0.0	0.22 $\pm$ 0.0	61.25 $\pm$ 3.6
Discrete	93.68 $\pm$ 0.8	1.42 $\pm$ 0.0	43.02 $\pm$ 1.7	67.36 $\pm$ 4.8	1.24 $\pm$ 0.0	53.72 $\pm$ 4.9
Continuous (Ours)	97.15 $\pm$ 4.5	1.44 $\pm$ 0.1	28.70 $\pm$ 0.2	71.37 $\pm$ 2.1	1.29 $\pm$ 0.1	34.45 $\pm$ 2.4

Table 41: Performance comparison of scoring strategies on energy sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	WMT			KO		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	31.90 $\pm$ 2.5	1.24 $\pm$ 0.1	28.51 $\pm$ 1.6	6.53 $\pm$ 0.4	0.35 $\pm$ 0.0	13.66 $\pm$ 0.8
Discrete	135.94 $\pm$ 11.6	3.90 $\pm$ 0.0	10.74 $\pm$ 0.8	78.61 $\pm$ 4.1	2.60 $\pm$ 0.0	8.43 $\pm$ 0.4
Continuous (Ours)	140.92 $\pm$ 2.6	4.00 $\pm$ 0.2	8.64 $\pm$ 0.2	81.95 $\pm$ 4.8	3.86 $\pm$ 0.1	6.86 $\pm$ 1.3
	LAUR*			LRN*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	45.98 $\pm$ 1.7	1.35 $\pm$ 0.1	24.60 $\pm$ 2.3	-17.84 $\pm$ 0.6	-0.07 $\pm$ 0.0	73.39 $\pm$ 3.8
Discrete	226.23 $\pm$ 4.7	4.16 $\pm$ 0.2	11.54 $\pm$ 0.5	85.46 $\pm$ 5.4	1.37 $\pm$ 0.1	36.78 $\pm$ 2.1
Continuous (Ours)	229.92 $\pm$ 3.4	4.20 $\pm$ 0.2	9.13 $\pm$ 0.2	87.39 $\pm$ 0.1	1.39 $\pm$ 0.1	30.72 $\pm$ 2.9

Table 42: Performance comparison of scoring strategies on consumer staples sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index, and bold values indicate the best performance.

	BTC-USD		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
Random	13.66 $\pm$ 0.2	0.47 $\pm$ 0.0	42.32 $\pm$ 0.6
Discrete	125.51 $\pm$ 6.0	1.87 $\pm$ 0.1	16.03 $\pm$ 1.6
Continuous (Ours)	128.36 $\pm$ 1.4	1.89 $\pm$ 0.1	15.97 $\pm$ 0.6

Table 43: Performance comparison of scoring strategies on BTC-USD ($p < 0.05$). Bold values indicate the best performance.

α	NVDA			AAPL		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
0	59.16 $\pm$ 2.4	1.15 $\pm$ 0.0	41.99 $\pm$ 0.6	53.45 $\pm$ 1.6	1.58 $\pm$ 0.1	24.37 $\pm$ 1.4
1	82.38 $\pm$ 8.1	1.19 $\pm$ 0.0	36.37 $\pm$ 2.8	65.07 $\pm$ 5.4	1.84 $\pm$ 0.2	20.62 $\pm$ 1.8
0.5	97.10 $\pm$ 7.2	1.57 $\pm$ 0.2	32.99 $\pm$ 1.7	67.76 $\pm$ 3.8	2.00 $\pm$ 0.0	17.07 $\pm$ 0.3
Learnable (Ours)	129.53 $\pm$ 4.5	1.89 $\pm$ 0.1	31.84 $\pm$ 3.4	86.19 $\pm$ 3.6	2.27 $\pm$ 0.1	9.35 $\pm$ 3.8
α	CIFR*			QBTS*		
	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$	AR% $\uparrow$	SR $\uparrow$	MDD% $\downarrow$
0	114.94 $\pm$ 11.3	1.22 $\pm$ 0.1	74.29 $\pm$ 6.1	149.12 $\pm$ 1.0	1.35 $\pm$ 0.1	81.74 $\pm$ 3.3
1	152.69 $\pm$ 14.6	1.37 $\pm$ 0.1	71.26 $\pm$ 2.4	243.31 $\pm$ 10.5	1.53 $\pm$ 0.1	76.97 $\pm$ 2.1
0.5	211.19 $\pm$ 17.5	1.59 $\pm$ 0.1	65.97 $\pm$ 5.1	305.53 $\pm$ 12.8	1.72 $\pm$ 0.1	54.11 $\pm$ 0.6
Learnable (Ours)	249.69 $\pm$ 2.7	1.67 $\pm$ 0.1	57.14 $\pm$ 5.0	346.06 $\pm$ 3.6	1.81 $\pm$ 0.0	50.66 $\pm$ 3.8

Table 44: Ablation study on the gating mechanism of NewsRank with different values of α on technology sector stocks ($p < 0.05$). '*' denotes mid-cap companies selected from the Russell 2000 index. Bold values indicate the best performance.

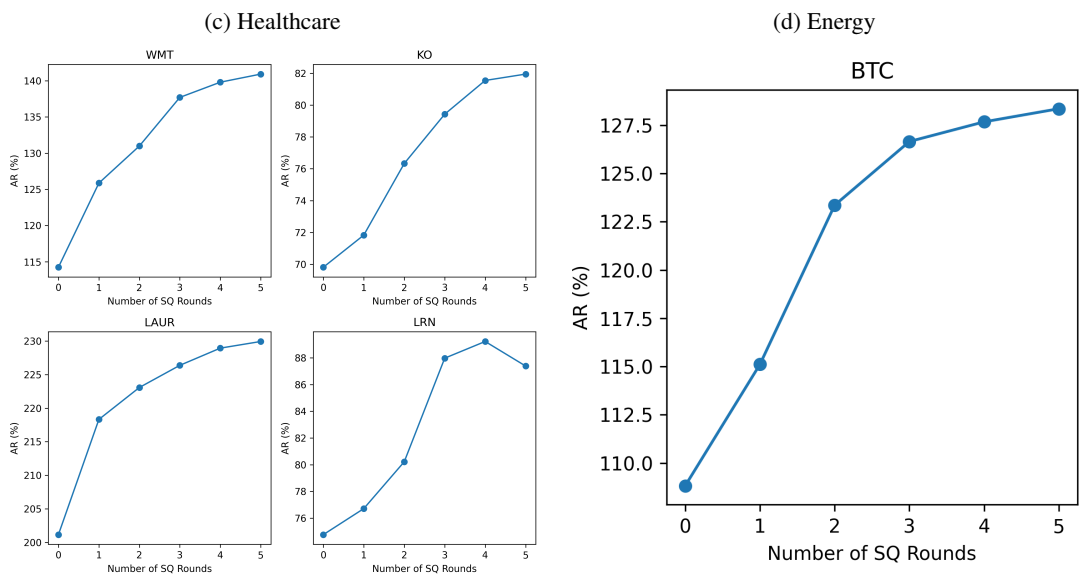
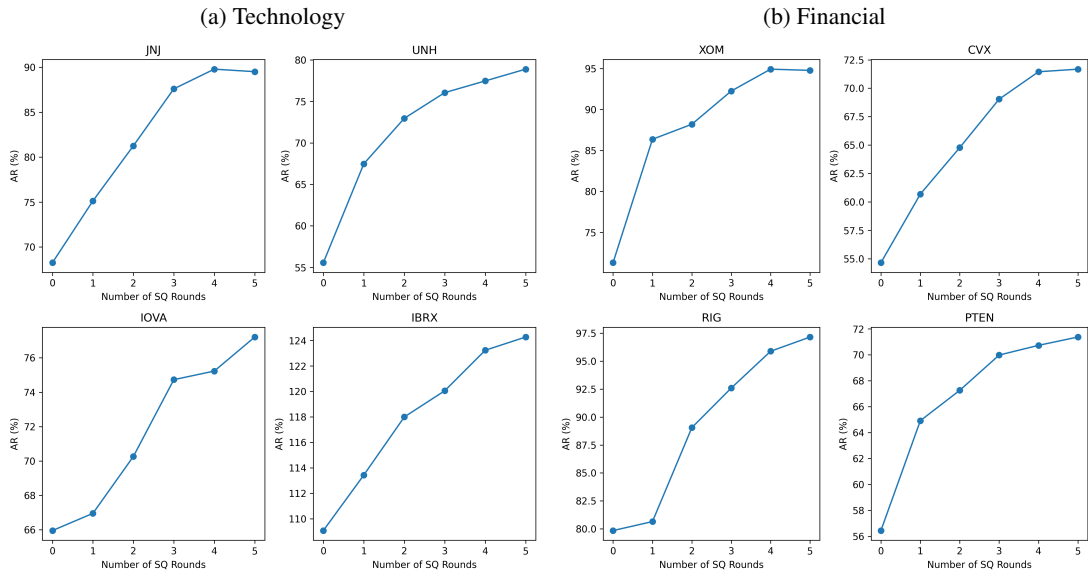
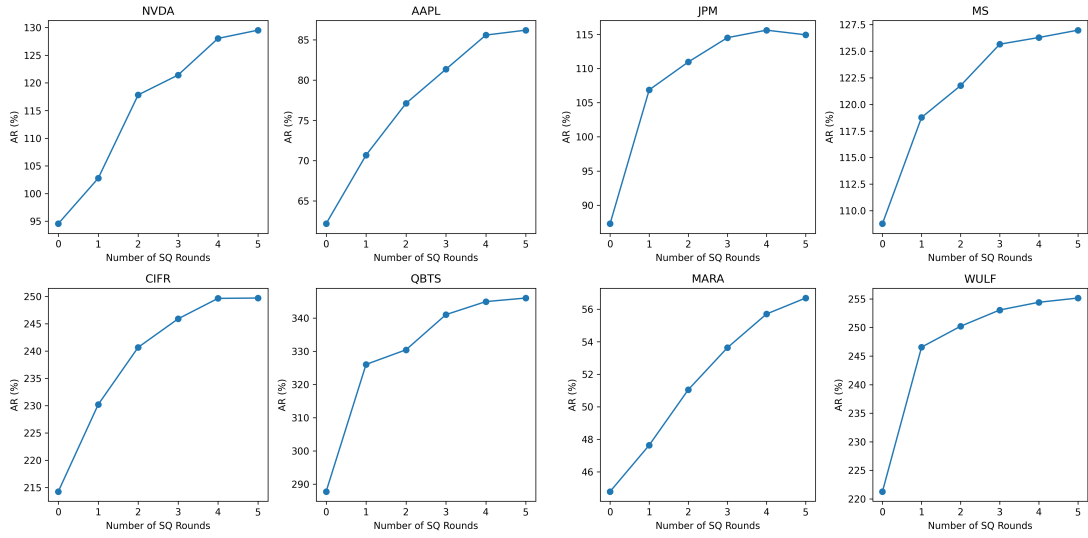
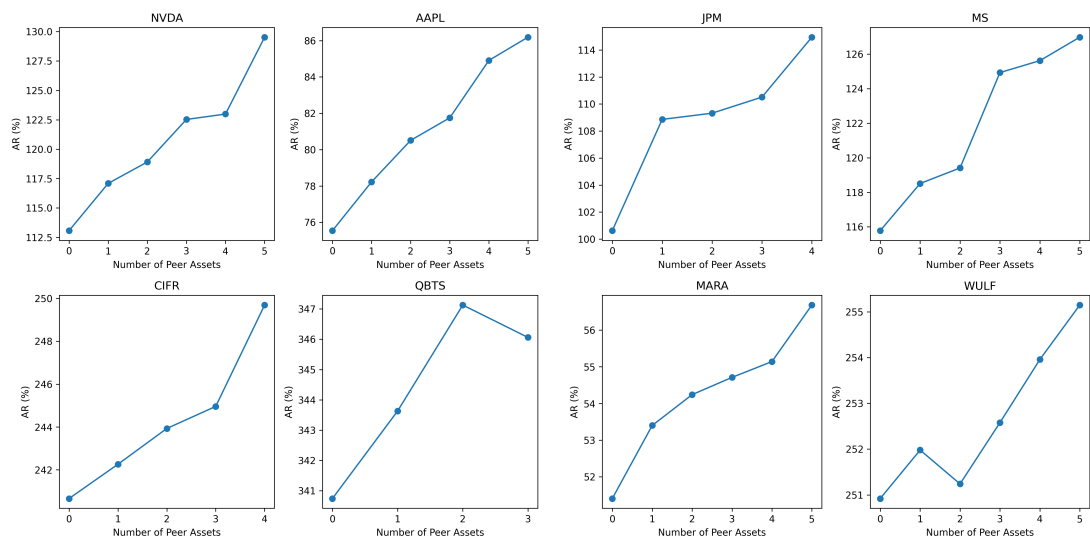
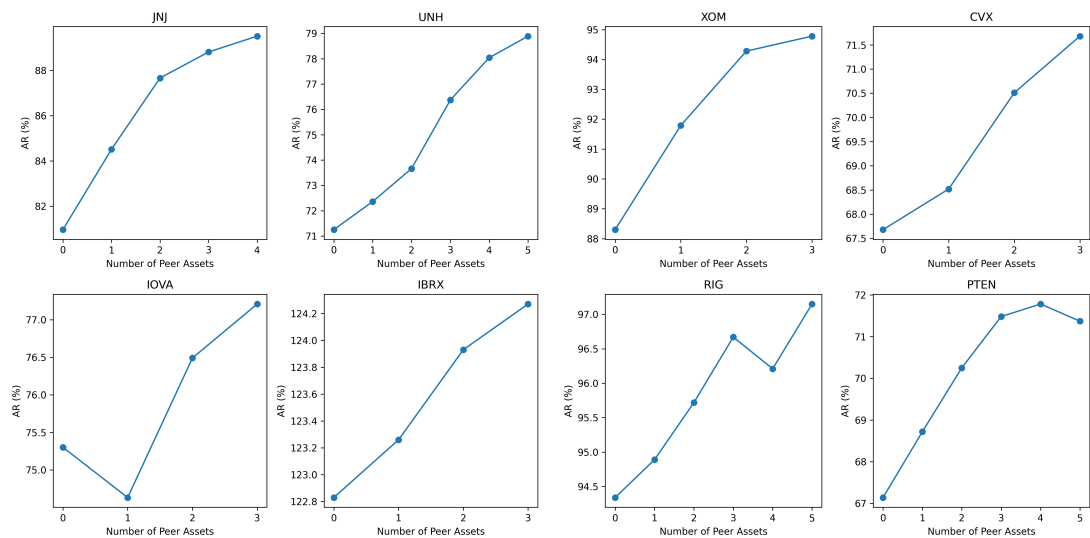


Figure 15: AR vs. Number of SQ Rounds Across Assets.



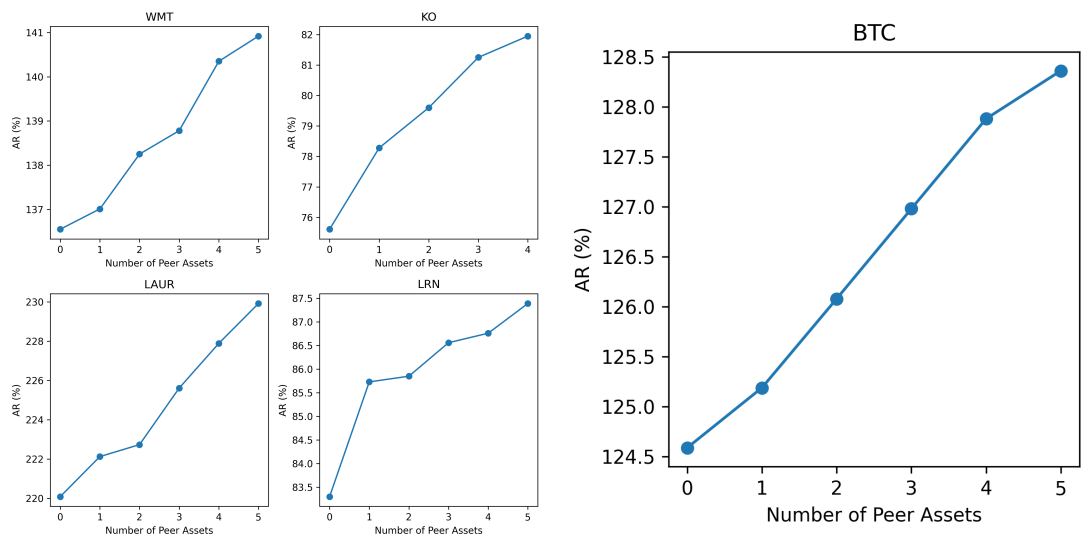
(a) Technology

(b) Financial



(c) Healthcare

(d) Energy



(e) Consumer

(f) BTC

Figure 16: AR vs. Number of Peer Assets Across Assets.

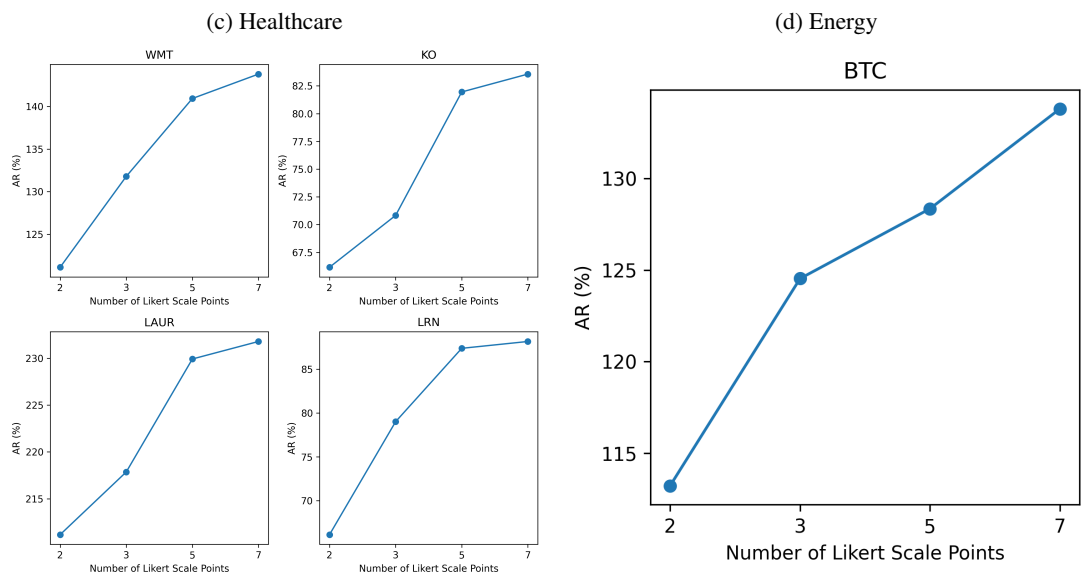
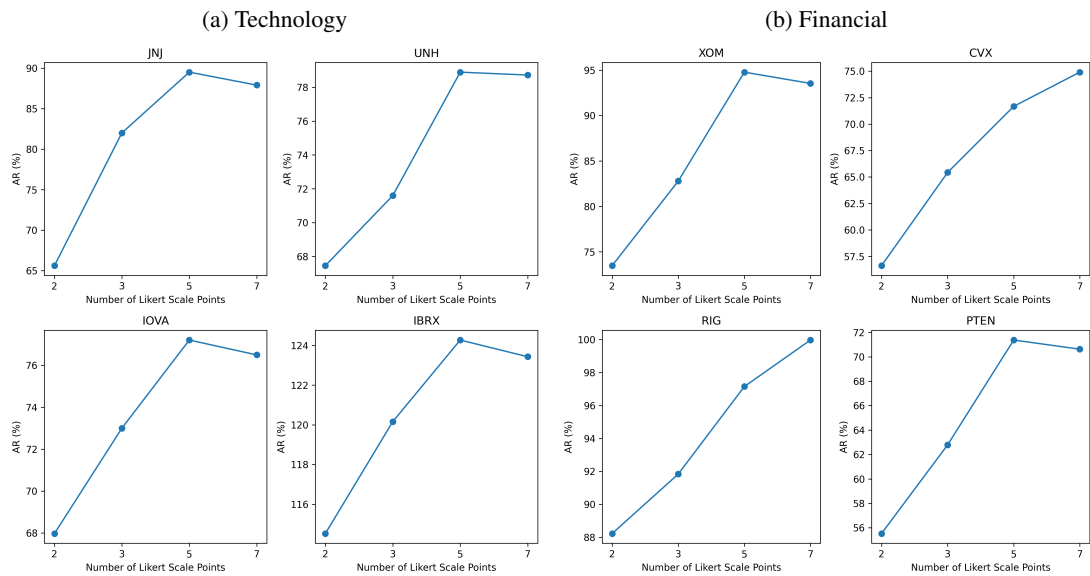
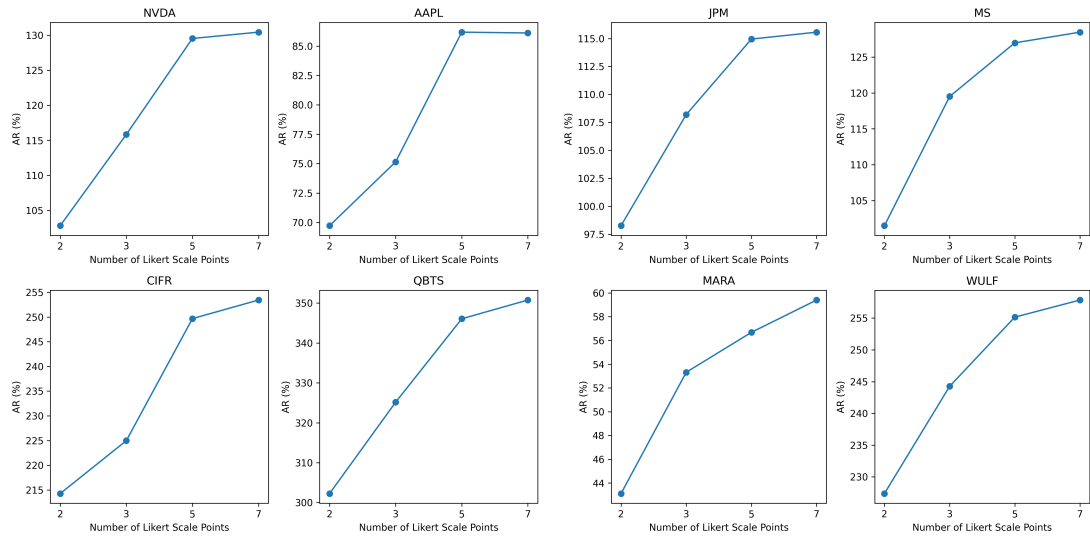


Figure 17: AR vs. Number of Likert Scale Points.

	Inference Latency (mins)	Number of LLM Calls	Input Tokens
LLMFactor	3.3 $\pm$ 0.2	3 $\pm$ 0.0	3,682 $\pm$ 143
FinMem	4.7 $\pm$ 0.3	5 $\pm$ 0.0	6,481 $\pm$ 167
FinAgent	5.4 $\pm$ 0.6	6 $\pm$ 0.0	12,746 $\pm$ 543
FinCon	9.6 $\pm$ 1.2	7 $\pm$ 0.0	14,468 $\pm$ 301
TA	13.7 $\pm$ 2.3	11 $\pm$ 0.0	16,482 $\pm$ 585
TiMi	9.4 $\pm$ 1.6	8 $\pm$ 0.0	9,661 $\pm$ 292
FinRAQ	12.3 $\pm$ 0.5	7 $\pm$ 0.0	9,481 $\pm$ 258

Table 45: Average per-inference costs over 21 assets during the testing period.

Date: 2024/11/12 - News A: The UK's CMA has provisionally found that Apple's policies restrict competition ... (4.6)
Date: 2024/11/25 - News B: India's CCI has rejected Apple's request to halt an investigation ... (2.9)
Date: 2025/03/12 - News C: The UK's CMA is investigating whether Apple and Google hold strategic market status ... (3.1 $\rightarrow$ 4.7) - News D: Apple has blocked rivals from accessing sensitive antitrust evidence in India, where the CCI investigation ... (4.0 $\rightarrow$ 2.4)

Figure 18: A case study of Bearish NewsRank refining news importance scores through semantic interactions. News articles with semantic interactions are highlighted in the same color. The refined importance scores are in bold, while the original importance scores are shown before the arrow.