RICE-ALPHA: RELIABILITY-INFORMED CORRECTION WITH EVENT GRAPHS FOR LLM-AGENT STOCK FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Equity-relevant news evolves through temporally dependent corporate events rather than isolated observations, making historical information useful only when event continuity, information availability, and transition reliability are modeled explicitly. Existing LLM-based financial agents increasingly incorporate historical evidence, yet they provide limited support for preserving issuer-specific chronology under point-in-time constraints and for identifying when historical transitions contribute information beyond the current forecast. We present RICE-ALPHA (Reliability-Informed Correction with Event Graphs), a point-in-time stock-scoring framework that separates a history-aware multi-view Base Alpha from a reliability-calibrated residual correction derived from historical event continuation. A Multi-Tier Memory Layer grounds news interpretation in temporally eligible issuer-specific history, while a Typed Event Agent constructs event states whose successor relations are formed within issuers and pooled across firms only after valid local pairing. Matured transitions are calibrated by their empirical reliability, and the resulting graph signal is residualized against the Base Alpha and technical view to obtain the RICE Delta. On daily Nasdaq-100 and Hang Seng Index panels from 2024 to 2026, RICE-ALPHA achieves the strongest results among the evaluated LLM-based agents and momentum across four predictive and four portfolio-level metrics in both markets. Its ICIR more than doubles that of the strongest baseline, while net Sharpe ratios reach 1.656 and 1.725 in the U.S. and Hong Kong, respectively. U.S. ablations further show significant reductions in IC and RankIC after Holm adjustment when major components are removed. These results indicate that historical event continuation adds incremental information when it is temporally grounded, reliability-calibrated, and introduced as a residual correction to a multi-view forecast.

1 INTRODUCTION

Financial decision-making requires integrating heterogeneous information from firm-level news (Boudoukh et al., 2019), financial statements (Ou & Penman, 1989), market dynamics (Gârleanu & Pedersen, 2013), and macroeconomic conditions (Fama & French, 1989). Concurrent advances in large language model (LLM) architectures (Liu et al., 2024a; Achiam et al., 2023), training paradigms, and data scaling have substantially improved their ability to represent and integrate heterogeneous information, providing a natural foundation for financial reasoning. Recent studies leverage LLMs to extract predictive signals from time-sensitive financial text and combine them with structured market information (Xiong et al., 2025; Lopez-Lira & Tang, 2026). Financial information, however, evolves continuously: newly observed news may update an existing corporate event, interact with prior issuer-specific developments, or acquire a different interpretation under changing market conditions. Effective financial reasoning therefore requires not only cutoff-available evidence, but also a temporally valid representation of how that evidence relates to historical event dynamics.

LLM-based financial agents increasingly retrieve historical information to contextualize current evidence (Yu et al., 2025; Guo et al., 2026; Qian et al., 2026), while event-based forecasting methods exploit structured news relations and temporal dependencies for return prediction (Han et al., 2025; Li et al., 2024a; 2026). Historical context and event structure, however, remain difficult to use

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reliably under a strict point-in-time setting. Retrieved evidence must remain consistent with what was observable at the prediction cutoff, while event relations should preserve issuer-specific chronology rather than arise from spurious cross-stock pairing. Structured event models show lag-dependent, time-varying inter-stock relations (Li et al., 2024a; Liu et al., 2024b), suggesting that the predictive relevance of historical transitions varies with empirical support, temporal stability, market context, and association with subsequent returns. These observations expose three coupled challenges in history-aware financial reasoning: **limited historical contextualization**, **weak temporal grounding of event relations**, and **uncalibrated use of historical transitions**.

To address these challenges, we present RICE-ALPHA (*Reliability-Informed Correction with Event Graphs*), a point-in-time stock-scoring framework that separates a history-aware multi-view reference signal from the incremental contribution of historical event continuation. As illustrated in Figure 1, RICE-ALPHA follows three stages. First, *History-Aware Multi-View Base Alpha Generation* constructs a point-in-time reference signal through dedicated sentiment, technical, fundamental, and macroeconomic agents. A structured Multi-Tier Memory Layer (MML) grounds the Sentiment Agent in temporally eligible issuer-specific history, while the remaining agents capture complementary cutoff-available market views. The resulting stock-level signals are reconciled before bounded macro adjustment to obtain Base Alpha. Second, *Point-in-Time Event Modeling and Reliability-Weighted Correction* is performed by the RICE Event Engine. A Typed Event Agent converts cutoff-available news and prior active records into structured event states. Historical successor relations are formed strictly within each issuer and pooled across firms only after valid local pairing, yielding a frozen point-in-time typed-event graph. Current states retrieve eligible one-hop transitions whose contributions are calibrated by transition frequency, regime-relative lift, lag structure, matured successor-return association, and observational reliability. The resulting graph signal is residualized against the Base Alpha and technical view to form the RICE Delta. Finally, *Final Alpha Synthesis* combines Base Alpha and RICE Delta through a bounded residual composition.

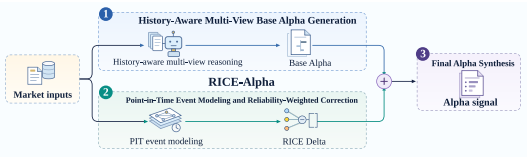


Figure 1: Overview of the RICE-Alpha framework. A point-in-time multi-view Base Alpha is refined by a reliability-calibrated residual derived from historical event propagation.

We evaluate RICE-ALPHA on daily U.S. and Hong Kong equity panels using four predictive metrics and four portfolio-level backtesting metrics. The empirical comparison focuses on LLM-agent stock-scoring systems and a price-based momentum baseline under a common evaluation protocol, while prior event-graph methods serve as methodological antecedents rather than directly benchmarked systems. Across both markets, RICE-ALPHA achieves the strongest results among the evaluated methods on all eight metrics. Ablation studies on the U.S. panel show significant reductions in IC and RankIC after Holm correction when major components are removed. Evaluations with an alternative LLM backbone, constituent-exclusion analysis, transaction-cost stress tests, and regime-specific analysis further examine robustness across modeling and trading conditions.

Our contributions are summarized as follows:

- **History-Aware Multi-View Base Alpha Generation.** We introduce a point-in-time multi-agent formulation that integrates sentiment, technical, fundamental, and macroeconomic evidence while grounding the Sentiment Agent in temporally eligible issuer-specific history through a structured MML. This enables historical context to inform current-news reasoning under a consistent point-in-time information set.
- **Point-in-Time Issuer-Local Event Graph Construction.** We introduce a Typed Event Agent that maps cutoff-available news to structured event records and typed states, and construct successor relations strictly within each issuer. Comparable transitions are pooled across issuers only after valid local pairing, preserving issuer-specific chronology while preventing look-ahead and spurious cross-stock transitions.
- **Reliability-Calibrated Residual Event Correction.** We develop a reliability-aware mechanism that converts matured one-hop transitions into a separate RICE Delta using transition frequency, regime-relative lift, lag structure, successor-return association, and observational reliability. The graph signal is residualized against the Base Alpha and technical view before final synthesis, isolating the incremental predictive content of historical event continuation.

2 Related Work

LLM-Based Agents for Financial Trading. Financial prediction has been studied through reinforcement learning and market representation (Wang et al., 2021; Liu et al., 2022; Li et al., 2024b), formulaic alpha discovery (Yu et al., 2023; Shi et al., 2025; Zhu & Zhu, 2025; Chen et al., 2026; Shi et al., 2026), and financial language models for textual reasoning and return forecasting (Araci, 2019; Wu et al., 2023; Yang et al., 2023; Xiao et al., 2024; Xie et al., 2023). Recent agentic systems further automate factor and strategy development (Zhang et al., 2026; Li et al., 2025; Song et al., 2026), while layered-memory and experience-retrieval mechanisms incorporate historical context into LLM-based decision making (Yu et al., 2025; Zhao et al., 2024; Fountas et al., 2025). Despite these advances, current market views and retrieved historical evidence are often processed through separate reasoning or signal-generation pathways (Yu et al., 2025; Xiao et al., 2024; Song et al., 2026), providing limited support for grounding current news in issuer-specific event history before cross-view aggregation. RICE-ALPHA differs by conditioning the sentiment view on temporally eligible historical evidence through a structured MML within a point-in-time multi-view Base Alpha.

Temporal Event Modeling and Reliability-Aware Forecasting. Structured event representations encode temporal, semantic, and causal dependencies in financial news (Han et al., 2025; Li et al., 2024a), while graph-based methods extend this formulation to cross-stock interactions, supply-chain relations, and dynamically evolving market structures (Qian et al., 2024; Huang et al., 2026; Yılkı, 2026; Zhang, 2026; Liu et al., 2024b). TRACE (Ding et al., 2026) combines temporal graph structure with mined rules and multi-hop reasoning, while retrieval- and memory-based systems incorporate historical availability, delayed feedback, and evolving contextual state (Zhao & Welsch, 2026; Deng et al., 2026; Yu et al., 2025; Fountas et al., 2025). These studies establish temporal event structure as an important source of predictive context. Relation construction over pooled observations, however, can attenuate issuer-specific chronology (Li et al., 2024a; Huang et al., 2026), whereas point-in-time forecasting requires queried relations, outcomes, and memory states to remain restricted to information available at the prediction cutoff (Zhao & Welsch, 2026; Deng et al., 2026). RICE-ALPHA differs in its construction principle by forming successor relations within each issuer and pooling comparable transitions only after valid local pairing under point-in-time constraints.

Graph-based stock models learn time-varying, lag-dependent relation strengths (Qian et al., 2024; Liu et al., 2024b; Li et al., 2024a), so historical transitions may differ in predictive relevance with empirical support, market regime, and return stability. Financial-agent frameworks increasingly decouple language-based reasoning from downstream numerical execution (Zhang et al., 2026; Li et al., 2025; Song et al., 2026; Koa et al., 2026), while structured consensus and temporal evaluation improve the stability of intermediate reasoning (Guo et al., 2026). These event-graph and event-reasoning approaches are methodologically related to RICE-ALPHA but are not directly benchmarked under the empirical protocol used in this study. Accordingly, our distinction is architectural rather than an empirical claim of superiority over prior event-graph methods. RICE-ALPHA represents matured transition evidence through a separate reliability-weighted residual, allowing its incremental contribution relative to the point-in-time Base Alpha to be isolated before final synthesis.

3 Methodology

3.1 Overview of RICE-Alpha

As illustrated in Figure 2, we present RICE-ALPHA (*Reliability-Informed Correction with Event Graphs*), a point-in-time stock-scoring framework that separates a history-aware multi-view reference signal from an explicit correction derived from historical event continuation. RICE-ALPHA comprises three modules: *History-Aware Multi-View Base Alpha Generation*, where dedicated sentiment, technical, fundamental, and macroeconomic agents process complementary financial views, with MML conditioning the Sentiment Agent on temporally eligible issuer-specific history before reconciliation into Base Alpha; *Alpha Correction via the RICE Event Engine*, which maps current and prior event records to typed states, queries a frozen PIT graph, and aggregates reliability-weighted one-hop transition evidence into RICE Delta; and *Final Alpha Synthesis*, which combines the two signals through a bounded residual composition. Cutoff-available news is therefore used along two complementary paths: MML supports history-aware semantic interpretation, while the Typed Event Agent and frozen event graph provide structured numerical evidence for event-continuation correction.

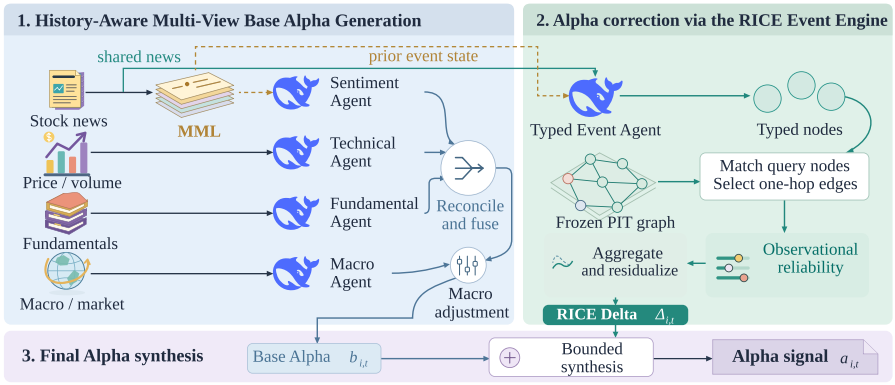


Figure 2: Overview of the RICE-ALPHA pipeline. Base Alpha captures multi-view evidence, RICE Delta captures reliability-weighted event continuation, and their bounded synthesis yields the final stock score.

At signal date t, given tradable universe $\mathcal{U}_t$ and cutoff-available news, price–volume history, fundamentals, and macro/market indicators, RICE-ALPHA outputs a cross-sectional stock-ranking score $a_{i,t}$ for each $i \in \mathcal{U}_t$. Evaluation follows a $T+1$ execution convention.

3.2 HISTORY-AWARE MULTI-VIEW BASE ALPHA GENERATION

For stock i at signal date t, the input state is $\mathcal{X}_{i,t} = (\mathcal{N}_{i,t}, \mathcal{T}_{i,t}, \mathcal{F}_{i,t}, \mathcal{M}_t)$, where $\mathcal{N}_{i,t}$ denotes cutoff-available stock news, $\mathcal{T}_{i,t}$ the price–volume history observed through t, $\mathcal{F}_{i,t}$ the admitted fundamental information, and $\mathcal{M}_t$ the dated macro/market indicators. The Sentiment, Technical, and Fundamental Agents process $\mathcal{N}_{i,t}$, $\mathcal{T}_{i,t}$, and $\mathcal{F}_{i,t}$, respectively, and return view-specific signals with associated confidence scores, while the Macro Agent maps $\mathcal{M}_t$ to a shared date-level market signal. Agent interfaces and validation rules are provided in Appendix A.

The Sentiment Agent is additionally conditioned on the Multi-Tier Memory Layer (MML), which retrieves temporally eligible issuer-specific context, including local graph context, matured historical cases, the current event narrative, regime-matched calibration evidence, and direct successor-event memories. Outcome-dependent memory is admitted only after the corresponding returns have matured. MML supports semantic interpretation of current news, whereas numerical transition statistics are handled separately by the RICE Event Engine in Section 3.3. Retrieval and memory-construction details are provided in Appendix Figure 5 and Appendix B.3.

Let $\mathbf{v}_{i,t} = (v^S_{i,t}, v^T_{i,t}, v^F_{i,t})$ denote the standardized sentiment, technical, and fundamental signals, and let $\widetilde{\mathbf{w}}_{i,t}$ denote their effective weights after confidence filtering, mature-feedback adjustment, and disagreement handling. Their weighted reconciliation defines $\widetilde{v}_{i,t}$, while the Macro Agent induces a stock-specific adjustment $m_{i,t}$. The Base Alpha is

$$b_{i,t} = \mathrm{clip}(m_{i,t}\widetilde{v}_{i,t}, b_{\min}, b_{\max}), \tag{1}$$

where $b_{\min}$ and $b_{\max}$ define the admissible Base Alpha range. Both $\widetilde{v}_{i,t}$ and $m_{i,t}$ are constructed exclusively from information available before the prediction cutoff. Detailed fusion, feedback, disagreement, macro-adjustment, and clipping definitions are provided in Appendix B. The resulting $b_{i,t}$ serves as the reference signal for the residual correction in Section 3.3.

3.3 ALPHA CORRECTION VIA THE RICE EVENT ENGINE

Conditioned on the Base Alpha $b_{i,t}$, the RICE Event Engine models predictive structure arising from the temporal evolution of corporate events. Issuer-specific news events form recurring successor patterns whose statistical regularities may persist across firms and market regimes. Semantic event extraction is performed by the Typed Event Agent, while graph construction, reliability weighting, aggregation, and residualization are deterministic. Figure 3 details its two-stage construction.

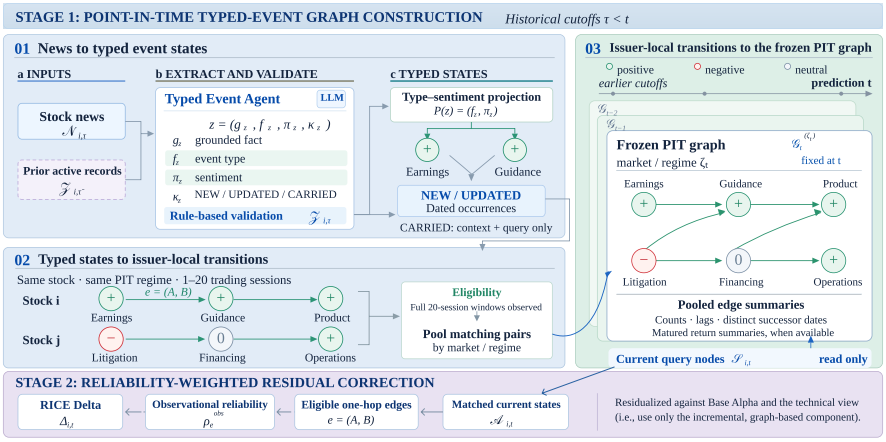


Figure 3: Overview of the RICE Event Engine. Stage 1 constructs a frozen point-in-time typed-event graph from issuer-local successor relations, while Stage 2 converts reliability-weighted graph evidence into the residual RICE Delta $\Delta_{i,t}$.

3.3.1 Stage 1: Point-in-Time Typed-Event Graph Construction

News to typed event states. Corporate news streams contain both genuine event progression and repeated reporting, so treating individual articles as independent events would distort transition statistics. At each historical cutoff $\tau < t$, the Typed Event Agent jointly processes stock news $\mathcal{N}_{i,\tau}$ and the previously validated active record set $\mathcal{Z}_{i,\tau^-}$, resolving newly reported information against existing issuer-specific episodes. Each accepted record is $z = (g_z, f_z, \pi_z, \kappa_z)$, where g_z is a grounded event description, f_z an event type, π_z an event-level sentiment label, and $\kappa_z \in \{\text{NEW}, \text{UPDATED}, \text{CARRIED}\}$ the lifecycle state. NEW opens an unmatched episode, UPDATED records substantive information about an active episode, and CARRIED preserves an unresolved episode without introducing a new occurrence. Only validated NEW and UPDATED records create dated graph occurrences, preventing repeated coverage from inflating transition support. Each occurrence is projected to $P(z) = (f_z, \pi_z)$, while episode identity and lifecycle metadata remain attached to the underlying record. At prediction cutoff t, typed states associated with current validated records, including CARRIED episodes, form the query set $\mathcal{S}_{i,t}$. Event taxonomy, episode matching, and graph-eligibility rules are specified in Appendix C.1.

Typed states to issuer-local transitions. Temporal succession is estimated within issuers to distinguish event evolution from incidental co-occurrence. For two eligible occurrences of the same stock with typed states A and B, Stage 1 admits $A \to B$ when B occurs within the successor window W_{succ} after A and both occurrences belong to the same market regime ζ. Issuer-local pairing preserves corporate-event chronology, while regime consistency separates successor patterns observed under distinct market conditions. Each directed pair represents temporal order rather than causation. An occurrence becomes an eligible anchor only after its full successor window has elapsed, preventing incomplete follow-up from being interpreted as absence of successors. The successor window and regime construction are specified in Appendices C.2 and B.3.

Issuer-local transitions to the frozen PIT graph. Individual firms typically provide too few repeated transitions for stable estimation, so Stage 1 pools comparable directed pairs across firms only after issuer-local formation and within the same market and regime ζ. The graph at cutoff t is $\mathcal{G}_t^{(\zeta)} = (\mathcal{V}_t^{(\zeta)}, \mathcal{E}_t^{(\zeta)})$, where $A = (f, \pi) \in \mathcal{V}_t^{(\zeta)}$ denotes a typed event state and $e = (A, B) \in \mathcal{E}_t^{(\zeta)}$ a matured successor relation. This local-first, pool-second construction combines issuer-specific chronology with cross-sectional statistical strength. Each pooled edge summarizes transition support, regime-relative frequency, lag structure, distinct successor dates, and matured successor-return behavior. The graph is frozen at cutoff t, so current records may query but cannot update the statistics used for their own prediction. Pooling rules, maturity conditions, and graph-construction hyperparameters are specified in Appendix C.2.

3.3.2 STAGE 2: RELIABILITY-WEIGHTED RESIDUAL CORRECTION

Current states to one-hop successor edges. At cutoff t, the current query set is matched against the frozen graph for regime ζ_t, yielding $\mathcal{A}_{i,t} = \mathcal{S}_{i,t} \cap \mathcal{V}_t^{(\zeta_t)}$. For each matched state A, Stage 2 considers eligible one-hop outgoing edges $e = (A, B)$, restricting propagation to directly observed successor structure. Let n_A denote the number of eligible anchors of A, n_e the number followed by B, $n_{\to B}$ the number of regime-matched anchors followed by B, and N_ζ the total number of eligible anchors in regime ζ. The smoothed conditional transition rate $p_e = (n_e + \alpha_p)/(n_A + \alpha_p + \beta_p)$ estimates how frequently B follows A, while $q_e = (n_{\to B} + \alpha_q)/(N_\zeta + \alpha_q + \beta_q)$ provides the corresponding regime baseline. Their ratio $L_e = p_e / \max(q_e, \varepsilon)$ defines relative lift, discounting broadly prevalent successors. Candidate edges are restricted to non-self transitions and filtered by support, transition strength, and per-node budget constraints, with hyperparameters specified in Appendix C.2.

Successor edges to reliability weights. Transition recurrence alone is insufficient to establish predictive relevance. Each retained edge therefore receives an observational reliability weight $\rho_e^{\text{obs}} \in [0, 1]$ based on matured transition support, coverage across distinct successor dates, lag behavior, and standardized successor-return association. Successor returns are admitted only after maturity, adjusted within market and date, and aggregated at the successor-date level to prevent repeated observations from dominating the estimate. Full estimators and reliability hyperparameters are provided in Appendices C.2 and C.3.

Weighted edges to the residual RICE Delta. Each retained edge $e = (A, B)$ contributes a bounded score $C_{i,e,t}$ whose sign follows the matured successor-return association and whose magnitude combines transition strength, observational reliability, query-state importance, and temporal alignment. Aggregating retained outgoing edges yields

$$r_{i,t}^{\text{graph}} = \text{clip}\left(\sum_{A \in \mathcal{A}_{i,t}} \sum_{e \in \mathcal{E}_t^{(\zeta_t),\to}(A)} C_{i,e,t}, -\tau_\Delta, \tau_\Delta \right), \tag{2}$$

where $\mathcal{E}_t^{(\zeta_t),\to}(A)$ denotes the retained outgoing edges of A and τ_Δ bounds the aggregate graph contribution. Because MML and the technical view may already encode information associated with the same underlying events, direct addition can duplicate information contained in the reference forecast. The graph signal is therefore residualized against the Base Alpha and technical view:

$$\Delta_{i,t} = g_{i,t}^{\text{edge}} \, \text{Res}_t\left(r_{\cdot,t}^{\text{graph}}; b_{\cdot,t}, v_{\cdot,t}^T \right)_i, \tag{3}$$

where $g_{i,t}^{\text{edge}} \in \{0, 1\}$ indicates whether stock i has at least one retained non-self transition. The gating restricts correction to observations supported by matched historical evidence, while Res_t removes cross-sectional rank-linear overlap with the controls before standardization. Edge-scoring coefficients, lag parameters, clipping bounds, residualization settings, and fallback rules are specified in Appendix C.3.

3.4 FINAL ALPHA SYNTHESIS

The RICE Delta $\Delta_{i,t}$ is combined with Base Alpha $b_{i,t}$ through a bounded residual composition:

$$a_{i,t} = \text{clip}(b_{i,t} + \lambda_\Delta s_\Delta \Delta_{i,t}, a_{\min}, a_{\max}), \tag{4}$$

where $s_\Delta > 0$ aligns the standardized residual correction with the Base Alpha scale, λ_Δ controls its contribution, and $a_{\min}$ and $a_{\max}$ define the admissible score range. Concrete synthesis parameters and score bounds are reported in Appendix Table 7.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETTINGS

Task, data, and baselines. On each signal date, every method scores the Nasdaq-100 (U.S.) or Hang Seng Index (Hong Kong) constituents against the five-session return from the next open, using

date-specific membership, adjusted prices, financials, news (English and Chinese in Hong Kong), and market indicators (Alpha Vantage, 2026; Tushare, 2026). Following a 2023 warm-up that initializes memory and event statistics, evaluation runs from 2024-01-02 to 2026-03-30 (562 U.S. and 551 Hong Kong signal dates). LLM-agent baselines, MEME (Guo et al., 2026), R&D-Agent-Quant (Li et al., 2025), and AI Hedge Fund (Singh, 2024), test whether RICE-ALPHA adds value beyond existing agentic systems; price-only 12–1 momentum tests it against a simple trend signal. All methods share each market's data, period, targets, and dated universe (Appendix D.2). RICE-ALPHA uses DeepSeek-V4-Flash (DeepSeek-AI, 2026) without fine-tuning.

Portfolio and frozen configuration. Every strategy in Table 1 follows one long-only rule, so strategies differ only in their scores. At each weekly rebalance, the portfolio is reset to equal weights on the top-scoring 10% of constituents, trades at the next open, and is held until the next rebalance. It pays trading costs of 4+4 bp per side in the U.S. and 10+4 bp in Hong Kong (commission plus slippage; Appendix D.3). Prompts, hyperparameter values, update rules, the portfolio rule, and trading costs were frozen at the end of the 2023 warm-up, with identical hyperparameters in the two markets. No setting was tuned, selected, or revised using 2024–2026 data. During evaluation, memory, event statistics, feedback, and view weights change only through the frozen update rules (Appendices D.1 and B.2–C.2).

Metrics and tests. IC and RankIC (daily cross-sectional correlations of scores with realized returns) measure ranking quality, ICIR and RankICIR day-to-day stability, and ARR, Sharpe, maximum drawdown (MDD), and Calmar ratio (CR) net tradable outcomes. Newey–West t-statistics (five lags) handle the overlapping five-session targets, and one-sided paired block bootstraps with Holm adjustment compare the complete system with the ablation variants and AI Hedge Fund (Appendix D.3).

4.2 MAIN RESULTS

Table 1: Main results on common data and dated universes. ICIR and RankICIR are not annualized; ARR and MDD (a positive loss) are in %. Bold: market best. All strategies use the long-only, equal-weighted top-10% weekly rule and costs of Section 4.1; NDX/HSI exclude costs. †/‡: lag-5 Newey–West $t > 1.96/2.58$.

	U.S. (Nasdaq-100)				Hong Kong (HSI)			
Panel A: Prediction quality Method	IC ↑	ICIR ↑	RankIC ↑	RankICIR ↑	IC ↑	ICIR ↑	RankIC ↑	RankICIR ↑
MEME	0.0216	0.1135	0.0254	0.1228	0.0013	0.0067	0.0156	0.0794
R&D-Agent-Quant	0.0188	0.0796	0.0177	0.0706	0.0011	0.0048	0.0125	0.0630
AI Hedge Fund	0.0077	0.0595	0.0094	0.0845	−0.0018	−0.0121	0.0033	0.0264
Momentum (12–1)	0.0316	0.1274	0.0322	0.1324	0.0160	0.0645	0.0327	0.1312
RICE-ALPHA (ours)	**0.0350**‡	**0.2928**	**0.0360**‡	**0.2927**	**0.0283**†	**0.1986**	**0.0395**‡	**0.2514**
Panel B: Strategy performance Method	ARR ↑	Sharpe ↑	MDD ↓	CR ↑	ARR ↑	Sharpe ↑	MDD ↓	CR ↑
Index (NDX / HSI)	20.66	0.986	24.37	0.848	22.24	0.952	21.24	1.047
MEME	10.60	0.591	22.38	0.474	11.07	0.638	15.72	0.704
R&D-Agent-Quant	11.62	0.567	27.24	0.427	17.43	0.929	16.32	1.068
AI Hedge Fund	10.97	0.700	17.52	0.626	27.84	1.130	20.13	1.383
Momentum (12–1)	18.94	0.889	21.46	0.883	23.10	1.098	13.23	1.746
RICE-ALPHA (ours)	**30.86**	**1.656**	**12.09**	**2.552**	**36.64**	**1.725**	**11.83**	**3.097**

RICE-ALPHA ranks first on every metric in both markets. As shown in Table 1, RICE-ALPHA supplies the strongest ranking signal in both markets, with the most notable improvement being its repeatability. Compared to momentum (the strongest baseline), IC rises from 0.0316 to 0.0350 in the U.S. and from 0.0160 to 0.0283 in Hong Kong, whereas the best agent baselines reach at most 0.0216 and 0.0013. ICIR more than doubles in both markets (0.2928 vs. 0.1274; 0.1986 vs. 0.0645), largely because the daily U.S. IC of RICE-ALPHA exhibits half the dispersion (SD 0.120 vs. 0.248), consistent with the event correction turning sparse historical episodes into a steadier daily ranking.

The advantage carries over to Hong Kong. Hong Kong is a demanding stress test: with the same frozen hyperparameters, it changes the language mix, disclosure sources, and market microstructure. MEME, R&D-Agent-Quant, and AI Hedge Fund degrade severely, reaching Hong Kong ICs of

0.0013, 0.0011, and −0.0018 (at most 6% of their U.S. values), while momentum drops from 0.0316 to 0.0160. In contrast, RICE-ALPHA retains 81% of its U.S. IC, and its Hong Kong RankIC (0.0395) exceeds its U.S. value (0.0360), the only LLM-based method whose signal stays largely intact across markets.

Statistical significance. Mean IC and RankIC are significantly positive in both markets (Newey–West t of 4.31 and 4.24 in the U.S., 2.48 and 3.10 in Hong Kong), whereas those of momentum and AI Hedge Fund are not significant in either market, and paired tests against AI Hedge Fund yield Holm-adjusted $p \leq 0.0096$ (Appendix Table 9).

Better rankings translate into superior portfolios. RICE-ALPHA achieves the highest ARR, Sharpe, and CR and the smallest MDD in each market, with Sharpe ratios of 1.656 in the U.S. and 1.725 in Hong Kong (best baselines: 0.889 and 1.130). Its net ARR exceeds the indices by 10.20 and 14.40 points with roughly half the drawdown (12.09% vs. 24.37%; 11.83% vs. 21.24%). Figure 4 shows how this edge accumulates: the Hong Kong strategy leads all baselines from late September 2024, while the U.S. strategy limits losses in the 2025 sell-off and pulls ahead in early 2026.

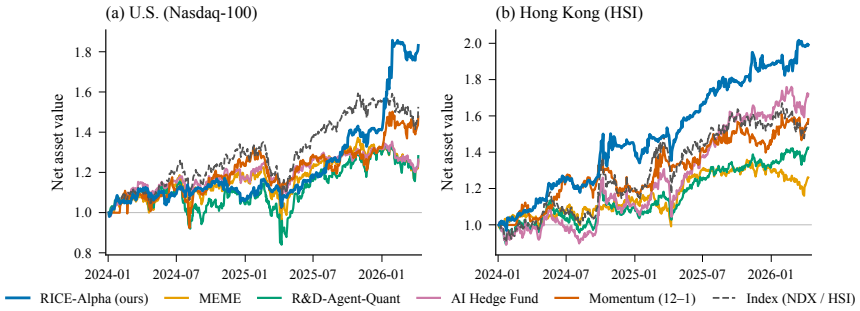


Figure 4: Net asset value of RICE-ALPHA, the baselines, and the index in (a) the U.S. and (b) Hong Kong from 2024-01-03; models are net of costs, indices exclude costs.

4.3 ABLATION STUDY

Table 2 introduces the components cumulatively on the U.S. panel, mapping each to a challenge of Section 1: retrieved memory (C1) addresses limited historical contextualization; numerical evidence from the issuer-local event graph (C2) resolves weak temporal grounding; and reliability weighting with residualization (C3) mitigates uncalibrated transitions. The *w/o C3* row injects the graph evidence as a direct correction without either C3 operation, two rows retain exactly one operation each to disentangle reliability weighting from residualization, and a separate run replaces the specialized sentiment, technical, and fundamental agents with one generalist stock-analysis agent (Appendix D.2).

Configuration	IC	ICIR	RankIC	RankICIR
RICE-ALPHA (complete)	**0.0350**	**0.2928**	**0.0360**	**0.2927**
w/o C3 (direct)	0.0260	0.2203	0.0290	0.2437
Reliability only	0.0326	0.2706	0.0330	0.2665
Residualization only	0.0208	0.1777	0.0242	0.2040
w/o C2–C3 (memory)	0.0191	0.1469	0.0205	0.1531
w/o C1–C2–C3	0.0064	0.0472	0.0090	0.0641
Single-agent run	0.0148	0.0925	0.0142	0.0828
Qwen3-Max backbone	0.0281	0.1958	0.0291	0.2012

Table 2: U.S. ablations and two separate references. C1 is retrieved memory, C2 numerical graph evidence, and C3 reliability weighting plus residualization. All component rows use multi-agent stock scoring; the single-agent row reports that run's stored final score. Component paired tests are in Table 9; the single-agent test is in the text.

Every component adds measurable signal. Both ranking quality and stability improve monotonically as components are added: IC rises from 0.0064 (no components) to 0.0191 (memory), 0.0260 (direct correction), and 0.0350 (full correction), and ICIR climbs from 0.0472 to 0.1469, 0.2203, and 0.2928, with RankIC and RankICIR following the same trajectory. All five ablation variants trail the complete system on IC and RankIC under six-row Holm tests (maximum adjusted $p = 0.0480$). Even with memory alone, the ICIR (0.1469) exceeds every baseline.

Reliability and residualization play complementary roles. Reliability weighting alone achieves an IC of 0.0326 (ICIR 0.2706), already surpassing momentum, and exceeds residualization alone (0.0208) with unadjusted one-sided $p = 0.0050$ for IC and 0.0406 for RankIC (Table 9, Panel C). Adding residualization on top of reliability weighting raises IC further to 0.0350 and ICIR to 0.2928, a Holm-significant gain. Reliability weighting thus acts as the primary safeguard against noisy continuations, and residualization adds value on top of it by excising overlap with the Base Alpha.

Multi-agent versus single-agent scoring. The single-agent run reaches IC 0.0148 and RankIC 0.0142. On the common U.S. panel, the complete system exceeds it by 0.0202 in IC and 0.0218 in RankIC (two-sided block bootstrap, $p = 0.020$ and 0.042; Appendix D.3), so the comparison favors the complete multi-agent configuration over the evaluated single-agent implementation.

Robustness across backbone, size, and cost. With Qwen3-Max in every LLM role, U.S. IC is 0.0281 and ICIR 0.1958, above every agent baseline (at most 0.0216 and 0.1135). Dropping ten large caps keeps IC and RankIC significant in both markets, and per-side costs of 20 and 28 bp leave Sharpe ratios of 1.551 and 1.659, above every default-cost baseline (Appendices E.2 and E.4).

4.4 Behavior in Falling Markets

Table 3 compares all methods on index-down and index-up sessions (Appendix E.5). Because all strategies share one portfolio rule, gaps between them reflect stock selection, whereas the gap to the index partly comes from the rule itself.

Table 3: Behavior in falling and rising markets. Capture: mean return on index-down/up sessions relative to the index. Crash: return (%) in the index's largest drawdown. Down mo.: mean return (%) in the 7 and 13 months the index fell. Bold: best Crash and Down mo.

	U.S. (Nasdaq-100)			Hong Kong (HSI)		
Method	Capture	Crash	Down mo.	Capture	Crash	Down mo.
Index (NDX / HSI)	1.00/1.00	−24.4	−4.5	1.00/1.00	−21.2	−3.3
MEME	0.28/0.33	−20.8	−4.1	0.02/0.11	**+3.5**	+0.2
R&D-Agent-Quant	0.29/0.35	−24.9	−4.9	0.25/0.33	−11.8	−1.8
AI Hedge Fund	0.73/0.71	−17.0	−3.2	0.93/0.97	−20.1	−2.8
Momentum (12–1)	0.83/0.85	−16.8	−4.1	0.58/0.64	−9.3	−1.4
RICE-ALPHA (ours)	0.51/0.65	**−10.9**	**−0.4**	0.34/0.51	−9.2	**+1.3**

RICE-ALPHA cuts losses more than gains. On index-down sessions, RICE-ALPHA captures 51% (U.S.) and 34% (Hong Kong) of the index decline, versus 83% and 58% for momentum and 73% and 93% for AI Hedge Fund, while retaining 65% and 51% of up-session gains. MEME and R&D-Agent-Quant lose less on ordinary down sessions owing to low market exposure (U.S. beta 0.20 and 0.18), but capture only a third of rallies and still lost 20.8% and 24.9% in the Nasdaq-100's largest crash (index 24.4%), whereas RICE-ALPHA lost 10.9%, the least of all methods. In months when the index fell, RICE-ALPHA posted the best mean return in both markets (-0.4% and $+1.3\%$) and outperformed the index in 6 of 7 U.S. months and all 13 Hong Kong months.

5 Conclusion

RICE-ALPHA addresses how historical corporate-event dynamics can improve stock forecasting without violating point-in-time constraints or duplicating information already captured by the current forecast. It combines temporally eligible issuer history for Base Alpha generation with an issuer-local event graph that calibrates matured transitions by reliability and residualizes their contribution into RICE Delta. Across the Nasdaq-100 and Hang Seng Index panels, RICE-ALPHA achieves the strongest results among the evaluated methods, while ablations confirm the contribution of memory, graph evidence, reliability weighting, and residualization. These findings remain robust across alternative backbones, constituent exclusions, higher transaction costs, and falling-market conditions, but are limited to two large-cap universes and pretrained LLM backbones whose training data may postdate individual forecasts. Overall, the results support point-in-time, issuer-local, reliability-calibrated event continuation as an effective source of incremental predictive information.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alpha Vantage. API documentation, 2026. URL https://www.alphavantage.co/documentation/. Accessed September 26, 2026.

Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019. URL https://arxiv.org/abs/1908.10063.

Jacob Boudoukh, Ronen Feldman, Shimon Kogan, and Matthew Richardson. Information, trading, and volatility: Evidence from firm-specific news. *The Review of Financial Studies*, 32(3): 992–1033, 2019. doi: 10.1093/rfs/hhy083.

Binqi Chen, Hongjun Ding, Ning Shen, Taian Guo, Jinsheng Huang, Luchen Liu, and Ming Zhang. AlphaSAGE: Structure-aware alpha mining via GFlowNets for robust exploration. In *International Conference on Learning Representations*, pp. 2955–2982, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/06115011841a8d500050261c0ac4b1e4-Abstract-Conference.html.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. arXiv preprint arXiv:2606.19348, 2026. URL https://arxiv.org/abs/2606.19348.

Zihao Deng, Yining Zhu, Leiming Wang, Jingfei Lu, Junbo Wang, Chuncheng Ran, Yu Yang, Dixuan Yang, and Jikun Shen. FinEvolveBench: A benchmark for self-evolving agents on low-repetition tasks with implicit rewards. arXiv preprint arXiv:2606.06960, 2026. URL https://arxiv.org/abs/2606.06960v3. Version 3, revised 12 August 2026.

Qianggang Ding, Haochen Shi, Luis Castejón Lozano, Miguel Conner, Juan Abia, Luis Gallego-Ledesma, Joshua Fellowes, Gerard Conangla Planes, Adam Elwood, and Bang Liu. TRACE: Temporal rule-anchored chain-of-evidence on knowledge graphs for interpretable stock movement prediction. arXiv preprint arXiv:2603.12500, 2026. URL https://arxiv.org/abs/2603.12500.

Eugene F. Fama and Kenneth R. French. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 25(1):23–49, 1989. doi: 10.1016/0304-405X(89)90095-0.

Zafeirios Fountas, Martin A. Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou Ammar, and Jun Wang. Human-inspired episodic memory for infinite context LLMs. In *International Conference on Learning Representations*, pp. 77230–77268, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c05144b635df16ac9bbf8246bbbd55ca-Abstract-Conference.html.

Nicolae Gârleanu and Lasse Heje Pedersen. Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, 68(6):2309–2340, 2013. doi: 10.1111/jofi.12080.

Taian Guo, Haiyang Shen, Junyu Luo, Zhongshi Xing, Hanchun Lian, Jinsheng Huang, Binqi Chen, Luchen Liu, Yun Ma, and Ming Zhang. MEME: Modeling the evolutionary modes of financial markets. arXiv preprint arXiv:2602.11918, 2026. URL https://arxiv.org/abs/2602.11918.

Bing Han, Gang Li, Dandan Qiao, and Mingxuan Zheng. Structured event representation and stock return predictability. SSRN Working Paper, 2025. URL https://ssrn.com/abstract=5952915.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Yikuan Huang, Zheqi Fan, Kaiqi Hu, and Yifan Ye. Cross-stock predictability via LLM-augmented semantic networks. arXiv preprint arXiv:2604.19476, 2026. URL https://arxiv.org/abs/2604.19476.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993. doi: 10.1111/j.1540-6261.1993.tb04702.x.

Kelvin J. L. Koa, Jan Chen, Yunshan Ma, Huanhuan Zheng, and Tat-Seng Chua. Reasoning on time-series for financial technical analysis. In *International Conference on Learning Representations*, pp. 64059–64079, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/68c1b9784d982fa5575bb45ace7bd64e-Abstract-Conference.html.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. doi: 10.1214/aos/1176347265.

Shuqi Li, Yuebo Sun, Yuxin Lin, Xin Gao, Shuo Shang, and Rui Yan. CausalStock: Deep end-to-end causal discovery for news-driven multi-stock movement prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47432–47454, 2024a. doi: 10.52202/079017-1504. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54d689d58fe54c92aee2d732fc49fca8-Abstract-Conference.html.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-guided stock transformer for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 162–170, 2024b. doi: 10.1609/aaai.v38i1.27767. URL https://ojs.aaai.org/index.php/AAAI/article/view/27767.

Yuante Li, Xu Yang, Xiao Yang, Xisen Wang, Weiqing Liu, and Jiang Bian. R&D-Agent-Quant: A multi-agent framework for data-centric factors and model joint optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131781–131830, 2025. doi: 10.52202/085713-3967. URL https://papers.neurips.cc/paper_files/paper/2025/hash/ac5c2b6e423883cbcacbcccf88491b78-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Zhixi Li, Jie Xiong, Jun Wang, Jinghua Tan, Philippe Du Jardin, Muhammet Deveci, and Kaiyang Zhong. A multi-graph learning framework to fuse heterogeneous market information for stock forecasting. *Expert Systems with Applications*, 319:132142, 2026. doi: 10.1016/j.eswa.2026.132142.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Mengpu Liu, Mengying Zhu, Xiuyuan Wang, Guofang Ma, Jianwei Yin, and Xiaolin Zheng. ECHO-GL: Earnings calls-driven heterogeneous graph learning for stock movement prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13972–13980, 2024b. doi: 10.1609/aaai.v38i12.29305.

Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Wang, Zhaoran Wang, and Jian Guo. FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1835–1849, 2022. doi: 10.52202/068431-0134. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0bf54b80686d2c4dc0808c2e98d430f7-Abstract.html. Datasets and Benchmarks Track.

Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT forecast stock price movements? Return predictability and large language models. *Journal of Financial Economics*, 184:104335, 2026. doi: 10.1016/j.jfineco.2026.104335.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610. URL https://doi.org/10.2307/1913610.

NVIDIA Corporation. NVIDIA announces financial results for fourth quarter and fiscal 2024. NVIDIA Newsroom press release, 2024. URL https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2024. Dated February 21, 2024; accessed August 16, 2026.

Jane A. Ou and Stephen H. Penman. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics*, 11(4):295–329, 1989. doi: 10.1016/0165-4101(89)90017-7.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. URL http://ilpubs.stanford.edu:8090/422/.

Hao Qian, Hongting Zhou, Qian Zhao, Hao Chen, Hongxiang Yao, Jingwei Wang, Ziqi Liu, Fei Yu, Zhiqiang Zhang, and Jun Zhou. MDGNN: Multi-relational dynamic graph neural network for comprehensive and dynamic stock investment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14642–14650, 2024. doi: 10.1609/aaai.v38i13.29381.

Xiaojun Qian, Sidi Chen, and Zhenqiang Sun. Sleipnir: Multi-agent based stock operator framework with interpretability. *IEEE Access*, 2026. doi: 10.1109/ACCESS.2026.3664216.

Hao Shi, Weili Song, Xinting Zhang, Jiahe Shi, Cuicui Luo, Xiang Ao, Hamid Arian, and Luis Angel Seco. AlphaForge: A framework to mine and dynamically combine formulaic alpha factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12524–12532, 2025. doi: 10.1609/aaai.v39i12.33365. URL https://ojs.aaai.org/index.php/AAAI/article/view/33365.

Yu Shi, Yitong Duan, and Jian Li. Navigating the alpha jungle: An LLM-powered MCTS framework for formulaic alpha factor mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 997–1005, 2026. doi: 10.1609/aaai.v40i2.37069. URL https://ojs.aaai.org/index.php/AAAI/article/view/37069.

Virat Singh. AI Hedge Fund. GitHub repository, 2024. URL https://github.com/virattt/ai-hedge-fund. Accessed September 26, 2026.

Zifan Song, Kaitao Song, Guosheng Hu, Ding Qi, Junyao Gao, Xiaohua Wang, Dongsheng Li, and Cairong Zhao. Trade in minutes! Rationality-driven agentic system for quantitative financial trading. In *International Conference on Learning Representations*, pp. 27140–27156, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2db9fbe4f34340cb928c653a6333084b-Abstract-Conference.html.

Tushare. Tushare Pro data API documentation, 2026. URL https://tushare.pro/document/2. Accessed September 26, 2026.

Zhicheng Wang, Biwei Huang, Shikui Tu, Kun Zhang, and Lei Xu. DeepTrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 643–650, 2021. doi: 10.1609/aaai.v35i1.16144. URL https://ojs.aaai.org/index.php/AAAI/article/view/16144.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564, 2023. URL https://arxiv.org/abs/2303.17564.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-agents LLM financial trading framework. arXiv preprint arXiv:2412.20138, 2024. URL https://arxiv.org/abs/2412.20138.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. The Wall Street neophyte: A zero-shot analysis of ChatGPT over multimodal stock movement prediction challenges. arXiv preprint arXiv:2304.05351, 2023. URL https://arxiv.org/abs/2304.05351.

Guojun Xiong, Zhiyang Deng, Keyi Wang, Yupeng Cao, Haohang Li, Yangyang Yu, Xueqing Peng, Mingquan Lin, Kaleb E. Smith, Xiao-Yang Liu, Jimin Huang, Sophia Ananiadou, and Qianqian Xie. FLAG-Trader: Fusion LLM-agent with gradient-based reinforcement learning for financial trading. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13921–13934, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.716. URL https://aclanthology.org/2025.findings-acl.716/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. arXiv preprint arXiv:2306.06031, 2023. URL https://arxiv.org/abs/2306.06031.

Asef Yılkı. Supply chain propagation of textual signals: LLM embeddings and cross-sectional return predictability. arXiv preprint arXiv:2606.29290, 2026. URL https://arxiv.org/abs/2606.29290.

Shuo Yu, Hongyan Xue, Xiang Ao, Feiyang Pan, Jia He, Dandan Tu, and Qing He. Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5476–5486, 2023. doi: 10.1145/3580305.3599831. URL https://doi.org/10.1145/3580305.3599831. Applied Data Science Track.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W. Suchow, Denghui Zhang, and Khaldoun Khashanah. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. *IEEE Transactions on Big Data*, 11(6):3443–3459, 2025. doi: 10.1109/TBDATA.2025.3593370. URL https://doi.org/10.1109/TBDATA.2025.3593370.

Lin Zhang. LLM-enhanced dynamic financial knowledge graphs for cross-entity signal propagation and alpha discovery. arXiv preprint arXiv:2607.10932, 2026. URL https://arxiv.org/abs/2607.10932.

Wentao Zhang, Mingxuan Zhao, Jincheng Gao, Jieshun You, Huaiyu Jia, Yilei Zhao, Bo An, and Shuo Sun. AlphaForgeBench: Benchmarking end-to-end trading strategy design with large language models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 10268–10277, 2026. doi: 10.1145/3770855.3817500.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

Zijie Zhao and Roy E. Welsch. Point-in-time financial RAG with frozen LLMs and market-feedback adaptive retrieval. arXiv preprint arXiv:2605.31201, 2026. URL https://arxiv.org/abs/2605.31201.

Zhoufan Zhu and Ke Zhu. AlphaQCM: Alpha discovery in finance with distributional reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80463–80479. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhu25ag.html.

A AGENT INTERFACES AND PROMPTS

All LLM roles use pretrained DeepSeek-V4-Flash (DeepSeek-AI, 2026) without task-specific fine-tuning; the backbone check substitutes Qwen3-Max (Yang et al., 2025) in every role. Across initialization and evaluation, the recorded U.S. and Hong Kong runs contain 327,937 and 363,562 successful LLM invocations. Table 4 lists the main symbols and the output of each role. Scores are ranking signals, not percentage-return forecasts.

Table 4: Main symbols and agent outputs. Only the five-session fields of the multi-horizon roles enter the evaluated score.

Object or role	Meaning or output
$a_{i,t}$, $b_{i,t}$, $\Delta_{i,t}$	Final score (Eq. 4), Base Alpha, and RICE Delta; $a^m_{i,t}$ is the evaluated score of method m
$o^j_{i,t}$, $v^j_{i,t}$, $c^j_{i,t}$	Raw offset, standardized score, and self-reported confidence of view $j \in \{S, T, F\}$
$z = (g_z, f_z, \pi_z, \kappa_z)$, $P(z)$	Typed-event record (fact, type, sentiment, lifecycle) and its graph node (f_z, π_z)
Episode / occurrence	Stock-local identity of a tracked event / validated, dated NEW or UPDATED record
$\mathcal{U}_t$, $\mathcal{U}^m_t$	Scoring universe on date t / its valid score–target subset for method m
Sentiment, Fundamental	`offsets`, `confidences`, and `reasonings` for 1, 5, and 20 sessions, plus `key_factors`
Technical	One 20-session `offset`, `confidence`, and `reasoning`, plus `key_factors`; also a residualization control
Macro	Offsets for the three windows; the five-session offset sets the shared macro signal
Typed Event	`daily_summary`, `active_events`, `resolved_events`, and `narrative_text` (Appendix C.1)

A scoring response is accepted only with a numeric offset in $[-50, 50]$, a confidence in $[0.05, 1]$, and nonempty reasoning; invalid responses are recorded separately from neutral answers. Accepted confidence is clipped to $[0.1, 1]$, and a view enters fusion only at confidence of at least 0.30 (Appendix B.1).

A.1 REQUEST TEMPLATES

The two system prompts below are reproduced verbatim; they are the entry points of current news into the sentiment and event-state paths. Angle-bracketed fields in the user messages stand for cutoff-available inputs. The Sentiment user message gives the security and its progressive narrative, followed by one section per available MML tier (in the order 0, 2, 1, 3, 4) and a provenance table with tier, source, freshness, similarity, and conflict flags. These section headings are legacy prompt labels (Tier 0, for example, is headed "Causal knowledge graph"); the graph records within-stock temporal order, not causal effects (Section 3.3). The Typed Event user message gives a covered-entity guard (entity identifier and aliases, with third-party companies marked as context only), the prior narrative with its date and update count, and the day's news items. The other scoring roles follow Table 4.

U.S. Sentiment – System Prompt

```
You are an event-driven analyst at a top quantitative hedge fund. Given a US stock's
progressive narrative (active / resolved / new / updated events plus narrative_text),
assess its impact over the next 1, 5, and 20 trading days.

Prioritize new and updated active events for 1d, broader digestion and institutional
follow-through for 5d, and only persistent structural drivers for 20d. Resolved or stale
background events should be sharply down-weighted.


## Multi-horizon output contract
Score the impact separately for **1d**, **5d**, and **20d**. Do not copy the
same value across horizons unless the evidence truly has the same expected
impact over all three holding periods.
```

U.S. Sentiment – System Prompt (continued)

```
-`1d`: immediate T+1 reaction and overnight/next-open repricing.
-`5d`: digestion over the next trading week, including analyst/fund-flow follow-through.
-`20d`: persistence, mean reversion, or structural trend over roughly one trading month.

Output JSON only:
{
"offsets": {"1d": 20, "5d": 12, "20d": -5},
"confidences": {"1d": 0.75, "5d": 0.65, "20d": 0.55},
"reasonings": {"1d": "...", "5d": "...", "20d": "..."},
"key_factors": ["...", "..."]
}
```

U.S. Typed Event – System Prompt

```
You are a senior US equity analyst maintaining
a point-in-time event state for exactly one covered stock.

Use only the prior state and news observable by the current date. Keep company events
that can still affect the next 5--20 trading days. Price action and technical indicators
are not events. A supplier, customer, peer, bank, analyst, or other mentioned company
must never replace the covered ticker as the subject.

Every active event is an object with exactly four fields:
-family: one of earnings, guidance_outlook, analyst_expectation,
clinical_regulatory, ma_divestiture, financing_capital_return,
litigation_investigation, regulation_policy_geopolitics, risk_incident_cyber,
governance_management_labor, partnership_contract, product_innovation,
operations_capacity_supply, pricing_demand_competition, esg_environment,
other_ineligible.
-polarity: exactly positive, negative, or mixed.
-change_type: exactly new, updated, or carried.
-grounded_text: one-line, verifiable event fact from the covered ticker's perspective;
it must not contain any RICE tag.

State-transition rules:
1. new means today's news created an event absent from the prior active state.
2. updated means today's news changed the factual status, strength, or polarity of one
prior event.
3. carried means a prior event remains active without a factual update today; its family
and polarity are unchanged.
4. Remove resolved/stale events from active_events and put a one-line fact in
resolved_events.
5. Never duplicate the same fact and never change family while an event remains active.
With no news, every retained event must be carried.
6. Earnings and one-day ratings usually expire after T+3; guidance and quarterly results
usually expire after 30 trading days; investigations, M&A, and major litigation can
remain active until clearly resolved.

Return strict JSON only, without extra fields or Markdown:
{
"daily_summary": "<=80 chars; with no news use: no new development; routine decay review",
"active_events": [
{"family":"earnings","polarity":"positive","change_type":"new","grounded_text":"..."}
],
"resolved_events": ["<one-line event facts newly resolved in this call>"],
"narrative_text": "<complete synthesized narrative, <=4000 chars>"
}

The covered entity and date are injected only by the runtime user prompt. Never infer or
substitute an entity from the system prompt.
```

B Base Alpha and Memory

B.1 Base Alpha

For stock i at cutoff t, let $v^j_{i,t}$, $c^j_{i,t}$, and $\widetilde{w}^j_{i,t}$ be the standardized score, confidence, and effective weight of view $j \in \{S, T, F\}$ (Appendix B.2), and let $\mathcal{J}_{i,t} = \{j : c^j_{i,t} \geq \tau_c\}$ be the eligible views.

Base Alpha in Eq. 1 is computed as

$$\bar{v}_{i,t} = \frac{\sum_{j\in\mathcal{J}_{i,t}} \widetilde{w}^{j}_{i,t} c^{j}_{i,t} v^{j}_{i,t}}{\sum_{j\in\mathcal{J}_{i,t}} \widetilde{w}^{j}_{i,t} c^{j}_{i,t}}, \qquad \widetilde{v}_{i,t} = \mathrm{clip}\big(s^{\mathrm{fb}}_{i,t}\,\mathrm{clip}(\bar{v}_{i,t}/\tau_v, -1, 1), -1, 1\big),$$
$$m_{i,t} = \mathrm{clip}\big(1 + \lambda_{\mathrm{macro}}\,\mu_t\,\chi_{i,t},\, m_{\min},\, m_{\max}\big), \qquad b_{i,t} = \mathrm{clip}\big(m_{i,t}\,\widetilde{v}_{i,t}, -1, 1\big), \tag{5}$$

with $\bar{v}_{i,t} = 0$ when the denominator is zero. Here $\mu_t = \mathrm{clip}(o^{\mathrm{Macro}}_t/\tau_\mu, -1, 1)$ is the Macro Agent's date-level signal, $\chi_{i,t} = \mathrm{clip}(\beta_{i,t}, \chi_{\min}, \chi_{\max})$ is the stock's historical market beta (χ_0 when unavailable), and $s^{\mathrm{fb}}_{i,t}$ is the matured-feedback scale (s^{fb}_0 without feedback). Extreme cross-view disagreement abstains with a zero score and zero confidence. Every adaptive quantity uses only information available, and outcomes matured, before the cutoff. Table 5 lists all settings.

Table 5: Base Alpha settings, frozen at the end of the 2023 warm-up. Initial weights are rounded; rolling weights and later adjustments remain adaptive only through these fixed update rules and cutoff-available information.

Setting	Value or rule
Eligible-view confidence τ_c; signal scale τ_v; macro-offset scale τ_μ	0.30; 3; 50
Macro strength λ_{macro}; bounds $(m_{\min}, m_{\max})$	0.25; (0.75, 1.25)
Beta bounds $(\chi_{\min}, \chi_{\max})$; missing-beta default χ_0; no-feedback scale s^{fb}_0	(0.3, 2.5); 1; 1
Initial EMA (mean, SD): S / T / F	(7, 15) / (3, 12) / (9, 12)
Prior strength n_0; EMA half-life H_{EMA}; score clip; SD threshold	10; 10; $[-3, 3]$; 0.01
Initial weights (S, T, F)	Approximately (0.180, 0.469, 0.351)
Rolling RankIC window / minimum	60 / 30 observations per view
Rolling weight score	$0.8\,Q + 0.2\,f_+$: Q is the rolling RankIC mean over its population SD (0 if the SD is $\le 10^{-6}$); f_+ is the share of positive RankIC
Downward-trend adjustment	Multiply by $\max(0.5, 1 + 10\,s)$ if the chronological slope $s < -0.001$
Weak-score floor	0.05 if $\lvert\overline{\mathrm{RankIC}}\rvert < 0.005$ or the score is nonpositive
Reconciliation gap boundaries; moderate confidence decrement	0.5 / 1.5 / 3.0; 0.10
High-gap weight / no-history scaling	0.5 / 0.5 before renormalization

B.2 Score normalization, reconciliation, and feedback

Normalization. For view j, let μ^j_t and σ^j_t be the cross-sectional mean and sample standard deviation of the n^j_t valid raw offsets. The lagged EMA states act as a shrinkage prior:

$$v^{j}_{i,t} = \mathrm{clip}\left(\frac{o^{j}_{i,t} - \mu^{\mathrm{eff},j}_t}{\sigma^{\mathrm{eff},j}_t}, -3, 3\right), \quad \mu^{\mathrm{eff},j}_t = (1-\nu^j_t)\mu^j_t + \nu^j_t \mu^{\mathrm{EMA},j}_{t-1}, \quad \nu^j_t = \frac{n_0}{n^j_t + n_0}, \tag{6}$$

and analogously for $\sigma^{\mathrm{eff},j}_t$. Scores use the old EMA states, which are updated afterwards with rate $\alpha = 1 - \exp(-\log 2/H_{\mathrm{EMA}})$; the SD state is updated only above its threshold. If $\sigma^{\mathrm{eff},j}_t \le 0.01$, all scores of view j are zero on that date; empty input leaves the state unchanged.

Reconciliation. Before fusion, the largest pairwise gap among a stock's normalized views sets its disagreement level (boundaries in Table 5). Moderate disagreement lowers the confidence scale; high disagreement halves the weights of views whose absolute cutoff-available regime RankIC is below half the largest one, after which weights are renormalized; extreme disagreement abstains. An abstention is a valid zero score, not a missing observation.

Matured-outcome feedback. Rolling view weights use matured daily RankIC across regimes; after a cold start, signal strength and sign consistency set normalized adaptive weights. Feedback

correlations are pooled Spearman correlations over valid historical stock–date records, conditioned on market regime, stock volatility, and momentum, with a fallback to the regime alone; support counts records, not dates. For composite correlation $I_{i,t}^{\mathrm{cmp}}$ with support $n_{i,t}^{\mathrm{cmp}}$, the Base Alpha scale is

$$s_{i,t}^{\mathrm{fb}} = \begin{cases} \mathrm{round}_2(1 + I_{i,t}^{\mathrm{cmp}}), & n_{i,t}^{\mathrm{cmp}} \geq 30,\ |I_{i,t}^{\mathrm{cmp}}| \geq 0.15, \\ 0.8, & n_{i,t}^{\mathrm{cmp}} \geq 30,\ |I_{i,t}^{\mathrm{cmp}}| \leq 0.05, \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

Separately, a view whose eligible regime RankIC is I receives the weight multiplier $\phi(I) = \min(1.2, 1 + I)$ for $I > 0.05$, 1 for $-0.02 < I \leq 0.05$, and $\max(0.7, 1 + 2I)$ for $I \leq -0.02$, rounded to two decimals; it equals one with fewer than 30 observations and is enabled for sentiment only. A separate confidence scale is $\mathrm{round}_2[\mathrm{clip}(0.4 + \bar{A}_{i,t}, 0.6, 1.2)]$, where $\bar{A}_{i,t}$ is the mean direction accuracy of supported views (one if none). Only outcomes matured before the cutoff are used, including in MML Tier 3. Adaptive weights use the population SD; the reported ICIR and RankICIR use the sample SD (Appendix D.3).

B.3 MULTI-TIER MEMORY LAYER AND MARKET REGIMES

MML supplies text only to the Sentiment Agent; the numerical correction is computed by the event engine. Table 6 lists the five retrieval units, and Figure 5 shows their shared filtering and assembly path. Graph-derived text and the numerical correction share the event history but use their own edge-selection rules.

Table 6: MML retrieval units and budgets. All entries must be available at the cutoff; outcome-based entries additionally require matured returns.

Tier	Context supplied	Selection budget
0	Typed-state neighborhood and communities	At most 500 candidate edges; 5 local edges and 3 communities
1	Similar historical cases with observable outcomes	At most 5 cases; similarity at least 0.4
2	Current narrative and active records	At most 1 entry
3	Regime-conditioned calibration context	At most 1 entry; at least 10 observable outcomes
4	Direct-successor summaries	At most 200 candidate edges; 3 outgoing summaries

Retrieval and text assembly. A historical case c is scored as $s_{\mathrm{case}}(c) = \xi_c(0.35s_{\mathrm{sem}} + 0.35s_{\mathrm{cat}} + 0.15s_{\mathrm{cont}} + 0.15s_{\mathrm{rec}})$, combining semantic, categorical, continuous-feature, and recency similarity, with $s_{\mathrm{rec}} = \exp[-\mathrm{clip}(\Delta_c^{\mathrm{td}}, 0, 252)/252]$ for the session distance Δ_c^{td} to the cutoff; without an embedding, the semantic weight moves to categorical similarity. Once a case's outcome matures, its multiplier becomes $\xi_c \leftarrow \mathrm{clip}(\xi_c e^{-0.1e_c}, 0.1, 2)$, where $e_c = 1$ for a direction mismatch with a nonzero stored score and $|r_c| \geq 0.0005$ for the filled realized return r_c, and $e_c = 0$ otherwise. Only filled cases at least 25 sessions old are eligible. Tier 0 edges and Tier 4 cards are ranked by $w_e^{\mathrm{ret}} = 0.15\log(1 + L_e p_e\sqrt{\max(n_e, 1)})$ (quantities in Appendix C.2) and require $p_e \geq 0.30$ and $L_e \geq 1.5$. Tier 0 searches a two-hop neighborhood and forms communities of at least two states by greedy modularity on the lift-weighted undirected graph; this structural text is separate from the one-hop numerical correction. Entries older than 180 sessions are excluded, the rendered text is capped at 6,000 characters, and opposing tier directions are kept and marked as a conflict when their gap reaches half the retrieved-entry count.

Market regimes. The date-level state ζ_t combines three volatility levels with risk-on, neutral, or risk-off conditions of the reference index (NDX in the U.S., HSI in Hong Kong), giving nine regimes. Volatility compares 20-session realized volatility with the 33rd and 67th percentiles of the available index history. Risk-on requires the 5-, 20-, and 60-session moving averages to decrease with lookback and the close to be at most 0.5% below the 5-session average; risk-off requires the reverse ordering and the close to be at most 0.5% above it. All other cases are neutral, and Mid/Neutral is used when

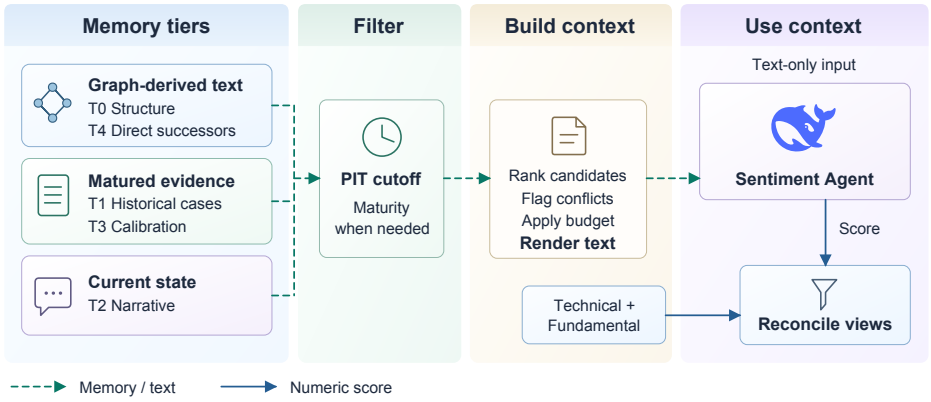


Figure 5: Multi-Tier Memory Layer. Graph-derived text (Tiers 0/4), matured cases and calibration (Tiers 1/3), and current narrative (Tier 2) pass through PIT and applicable maturity filters. Ranked, conflict-marked, budgeted text enters only the Sentiment Agent. Its numerical view joins Technical and Fundamental views for reconciliation before fusion. Dashed arrows carry memory/text; solid arrows carry numerical scores.

the index history is insufficient. Only prices dated through t enter. Each occurrence inherits its date's regime, and both ends of a within-stock pair must share it.

C TYPED-EVENT GRAPH AND RICE CORRECTION

This appendix gives the exact rules behind the two stages of Section 3.3; Table 7 collects their settings.

C.1 TYPED RECORDS AND LIFECYCLE

The Typed Event Agent (Appendix A.1) reads the company, the cutoff date, the prior narrative and active records, and the current news. Each active record has exactly four fields, `grounded_text`, `family`, `polarity`, and `change_type`, which correspond to $(g_z, f_z, \pi_z, \kappa_z)$; the sentiment *neutral*, which includes offsetting implications, is serialized as `mixed`. Rule-based checks then validate the lifecycle and assign or inherit a stock-local episode identity:

Lifecycle label	Episode rule	Dated occurrence
NEW	No prior active match; create a stock-local identity	Yes, after validation
UPDATED	Exactly one match; inherit its identity and event type	Yes, after validation
CARRIED	Keep the matched episode without a new fact	No; context and query only
Resolved	Close a prior episode omitted from the accepted active set	No

Sentiment can change through an UPDATED record but not a CARRIED one. Duplicate identity claims and ambiguous transitions are rejected; the prior records are then retained as context and no occurrence is created. Closed episodes are matched within stock and event type by normalized fact text, so a resolved fact cannot re-enter the active history, and records from different episodes may share a typed state without merging their histories. The validated set $\mathcal{Z}_{i,t}$ becomes the prior context at the next cutoff. The 15 graph-eligible types are earnings; guidance/outlook; analyst expectations; clinical/regulatory events; M&A/divestitures; financing/capital returns; litigation/investigations; regulation, policy, or geopolitics; risk incidents/cyber events; governance, management, or labor; partnerships/contracts; product innovation; operations/capacity/supply; pricing/demand/competition; and ESG/environmental events. Records of other types are ineligible. Sentiment states the cutoff-time implication for the stock, not a realized-return label. One release can yield several records: NVIDIA's February 21, 2024 report of \$22.1 billion quarterly revenue and a \$24.0 billion outlook ($\pm 2\%$) (NVIDIA Corporation, 2024) can be encoded as separate positive Earnings and Guidance records.

C.2 PAIRING, POOLING, AND EDGE ESTIMATES

Pairs and timing. Only validated NEW and UPDATED records form occurrences; every graph-eligible current record, CARRIED ones included, supplies a query node. An occurrence of state A is paired with each occurrence of state B of the same stock dated 1–20 exchange sessions later in the same regime, and pairs are pooled by market and regime only after this local pairing. The graph is re-estimated after every close. Each estimate reads only records dated at least 25 sessions before its date, and an occurrence anchors pairs only if its full 20-session window lies inside that range, so anchors are at least 45 sessions old. At cutoff t, Stage 2 reads the latest estimate of each edge dated no later than t, without an LLM call or a graph update; later estimates never alter the graph seen at an earlier date. The 20-session window governs graph construction only and does not change the five-session prediction target.

Counting units and transition rates. For an edge $e = (A, B)$ of the graph $\mathcal{G}_t^{(\zeta)}$, n_A counts A anchors, n_e the distinct anchors followed by at least one B, $n_{\to B}$ the anchors of the regime followed by B, and N_ζ all eligible anchors of the regime; $\mathcal{P}_e$ is the multiset of valid within-stock pairs, D_e the number of distinct successor dates, and $J_{e,h}$ the number of successor dates with usable matured h-session returns. With Beta$(1,1)$ priors ($\alpha_p = \beta_p = \alpha_q = \beta_q = 1$ and $\varepsilon = 10^{-12}$ in Section 3.3), the transition rate, regime baseline, lift, and lag profile are

$$p_e = \frac{n_e + 1}{n_A + 2}, \qquad q_e = \frac{n_{\to B} + 1}{N_\zeta + 2}, \qquad L_e = \frac{p_e}{\max(q_e, 10^{-12})},$$
$$\ell_e = \operatorname{mean}_{u \in \mathcal{P}_e} \operatorname{lag}(u), \qquad \sigma_{\ell,e} = \operatorname{pstdev}_{u \in \mathcal{P}_e} \operatorname{lag}(u). \tag{8}$$

Repeated B occurrences after one anchor count once in p_e but all enter the lag profile. Edges with $n_e \geq 5$, $L_e \geq 1.5$, and $p_e \geq 0.30$ are kept, at most 500 candidates per market–regime graph; each query node then keeps at most eight successors $B \neq A$ ranked by $p_e L_e$.

Post-successor returns. For $h \in \mathcal{H} = \{1, 5, 20\}$, the return of a successor on date d runs from the open of $d+1$ to the open of $d+h+1$ and is demeaned by the mean of all filled returns of the same market, date, and window, not only edge-linked stocks. Let $\bar{y}_{e,d,h}$ average these matured returns over the pairs of e that end on d, and let $\mathcal{D}^{Y}_{e,h}$ be the $J_{e,h}$ dates with at least one such return. Then

$$\theta_{e,h} = \frac{1}{J_{e,h}} \sum_{d \in \mathcal{D}^{Y}_{e,h}} \bar{y}_{e,d,h}, \qquad \widehat{\sigma}_{e,h} = \Big[\frac{1}{J_{e,h}} \sum_{d \in \mathcal{D}^{Y}_{e,h}} (\bar{y}_{e,d,h} - \theta_{e,h})^2\Big]^{1/2},$$
$$s^{\text{assoc}}_{e,h} = \theta_{e,h}\sqrt{J_{e,h}}/\widehat{\sigma}_{e,h}, \qquad \sigma^{\text{ret}}_{e,h} = \max(\widehat{\sigma}_{e,h}, \sigma_{\min}), \tag{9}$$

with $s^{\text{assoc}}_{e,h} = 0$ unless $J_{e,h} > 1$ and $\widehat{\sigma}_{e,h} > 10^{-12}$. When $J_{e,h} = 0$, $\theta_{e,h} = \widehat{\sigma}_{e,h} = s^{\text{assoc}}_{e,h} = 0$. Dates therefore receive equal weight regardless of how many pairs end on them; D_e counts all successor dates and can exceed $J_{e,h}$. The five-session values θ_e, s^{assoc}_e, and σ^{ret}_e enter the correction; the other windows inform reliability only.

C.3 Reliability, edge contribution, and residualization

Observational reliability. With the saturation thresholds of Table 7, the support product and the reliability weight are

$$R_e = \min\Big(1, \frac{n_e}{\tau_n}\Big) \min\Big(1, \frac{D_e}{\tau_D}\Big) \min\Big(1, \frac{n_e}{\tau_q}\Big) \min\Big(1, \frac{N_\zeta - n_A}{\tau_0}\Big),$$
$$\rho^{\text{stored}}_e = \text{clip}\Big[R_e \min\Big(1, \frac{\max_{h \in \mathcal{H}} |s^{\text{assoc}}_{e,h}|}{2}\Big), 0, 1\Big], \quad \rho^{\text{obs}}_e = \begin{cases} \rho^{\text{stored}}_e, & \rho^{\text{stored}}_e > 0, \\ R_e\, \psi(s^{\text{assoc}}_{e,5}), & \text{otherwise}, \end{cases} \tag{10}$$

where $\psi(x) = \min(1, |x|/2)$ for $0 < |x| < \infty$ and 0.5 otherwise. The factors of R_e measure sample size, successor-date coverage, a separate minimum for supporting anchors, and non-anchor support, that is, enough other anchors in the regime for the regime baseline q_e. Where the return statistic is undefined, ψ keeps half of the support weight, so the weight increases with consistency only where the statistic exists. The weight lies in $[0, 1]$ and is applied once, before residualization. It weights observed evidence; it is not a calibrated probability or a causal effect.

Edge contribution. The lag weight compares an edge's typical lag with the three reference windows, and the raw graph signal of Eq. 2 is formed with three clipping steps:

$$w^{\text{lag}}_{e,h} = \exp\Big[-\frac{1}{2}\Big(\frac{\ell_e - h}{\max(\sigma_{\ell,e}, 1.5)}\Big)^2\Big], \qquad \omega(e) = \frac{w^{\text{lag}}_{e,5}}{\sum_{h \in \mathcal{H}} w^{\text{lag}}_{e,h}},$$
$$\eta_e = \text{sclip}_{0.25}\big(\rho^{\text{obs}}_e \tanh(\log \widetilde{L}_e) \tanh(\theta_e/\sigma^{\text{ret}}_e)\big), \qquad C_{i,e,t} = \text{sclip}_{0.25}\big(x_{i,A,t}\, \gamma\, p_e\, \eta_e\, \omega(e)\big),$$
$$r^{\text{graph}}_{i,t} = \text{sclip}_{\tau_\Delta}\Big(\sum_{A \in \mathcal{A}_{i,t}} \sum_{e \in \mathcal{E}^{(\zeta_t),\rightarrow}_t(A)} C_{i,e,t}\Big), \tag{11}$$

where $\text{sclip}_a(x) = \min(a, \max(-a, x))$, $\widetilde{L}_e = \text{clip}(L_e, 10^{-6}, 20)$, γ is the depth coefficient, and $\tau_\Delta = 0.20$; no contribution is formed if the lag-weight denominator is numerically zero. Because kept edges have $L_e \geq 1.5$, the sign of $C_{i,e,t}$ is that of θ_e. The query-node weight $x_{i,A,t}$ is a personalized-PageRank score (Page et al., 1999) (damping 0.85) normalized to mean one over the stock's matched nodes, so a single matched node has weight one; equal weights are used if it is unavailable.

Residualization. The raw signal is first standardized across stocks with the sample SD and clipped at ± 3; the whole cross section is zero if that SD is at most 10^{-9}. Let R_t assign average ranks for ties and standardize them on each input's finite stocks. With $y_t = \mathsf{R}_t(z^\Delta)$ for the standardized signal z^Δ and the controls $\mathsf{R}_t(b)$ and $\mathsf{R}_t(v^T)$,

$$\widehat{\beta}_t = \arg\min_\beta \big\|y_{\mathcal{I}_t} - X_t\beta\big\|_2^2, \quad \varepsilon_t = y_{\mathcal{I}_t} - X_t\widehat{\beta}_t, \quad \mathsf{Res}_t(\cdot) = \frac{\varepsilon_t - \overline{\varepsilon}_t}{\text{sd}_t(\varepsilon_t)}, \tag{12}$$

where X_t holds an intercept and the nondegenerate controls on the common stock set $\mathcal{I}_t$, which must contain at least $\max(20, k+3)$ stocks for k controls. Without a usable control, y_t is returned directly. Degenerate responses or residuals give no correction; stocks outside $\mathcal{I}_t$ and stocks without a kept edge receive $\Delta_{i,t} = 0$. The residual is orthogonal to its controls before the edge mask and the final clip.

Daily order of operations. At each cutoff the system (1) restores the cutoff-available state; (2) validates current records and derives query nodes and occurrences without updating the queried graph; (3) assembles MML text for the Sentiment Agent; (4) obtains view scores; (5) normalizes them with the old EMA states, updates the states, reconciles the views, and forms Base Alpha; (6) reads the frozen edges and clips edge scores, contributions, and their sum in order; (7) standardizes, ranks, regresses on the controls, standardizes the residual, and applies the edge mask; and (8) adds the correction to Base Alpha and clips the final score. Later history affects only later cutoffs.

Table 7: Settings of the RICE Event Engine (Section 3.3) and of the final synthesis (Section 3.4). All listed parameter values and update rules were frozen at the end of the 2023 warm-up; graph estimates still update from cutoff-available records under those rules. Base Alpha settings are in Table 5.

Setting	Value
Stage 1: graph construction	
Within-stock successor window; record age at estimation	1–20 sessions; ≥ 25 sessions
Transition- and regime-rate priors; lift floor	Beta$(1,1)$; 10^{-12}
Kept edges: minimum support, lift, and transition rate	$n_e \geq 5$, $L_e \geq 1.5$, $p_e \geq 0.30$
Candidate edges per graph; successors per node; depth K	500; 8; 1
Stage 2: reliability, contribution, and residualization	
Saturation thresholds $(\tau_n, \tau_D, \tau_q, \tau_0)$	20, 8, 10, 30
Fallback ψ at a zero or unavailable statistic	0.5
Depth coefficient γ; lift domain	0.6; $[10^{-6}, 20]$
Return-scale floor $\sigma_{\min}$; lag centers; lag-dispersion floor	0.02; 1, 5, 20; 1.5 sessions
Edge-score and contribution clip; raw-signal bound τ_Δ	± 0.25; 0.20
Daily z-score clip; minimum common stocks with controls	± 3; 20
Final synthesis (Eq. 4)	
Correction coefficient λ_Δ; scale s_Δ; score bounds $(a_{\min}, a_{\max})$	0.30; $1/3$; $(-1, 1)$

D Data, Baselines, and Evaluation

D.1 Market coverage

Table 8 separates the initialization period from the prediction-evaluation period; backtest dates extend beyond the final signal to complete the last return window. Index membership is recorded for each date. In each market, all methods share the underlying data, evaluation interval, return targets, and dated universe. U.S. financial reports become available 90 calendar days after the fiscal period ends; Hong Kong uses the latest financial snapshot within a historical window that ends before the prediction date.

Warm-up and frozen configuration. The 2023 warm-up in Table 8 runs the daily pipeline to build MML memory, typed-event records, graph statistics, feedback statistics, EMA normalization states, and rolling view weights; its predictions enter no reported metric. At the end of the warm-up, the prompts (Appendix A), the parameter values and update rules of Tables 5 and 7, the ablation and backbone configurations of Appendix D.2, and the portfolio rule and trading costs of Appendix D.3 were frozen. Parameter values and update rules are identical in the two markets. No setting was tuned, selected, or revised using 2024–2026 data, and no part of that period serves as a validation set. During evaluation, the adaptive quantities above change only through the frozen update rules of Appendices B.2–C.2, using information available before each cutoff.

Table 8: Coverage of the two market panels. Valid pairs are RICE-ALPHA observations with finite scores and target returns.

Item	U.S. (Nasdaq-100)	Hong Kong (HSI)
Full run begins / final prediction	2023-01-03 / 2026-03-30	2023-01-03 / 2026-03-30
Full-run trading dates	812	794
Evaluation begins / signal dates	2024-01-02 / 562	2024-01-02 / 551
RICE-ALPHA valid stock–date pairs	56,746	46,345
Backtest period	2024-01-03 to 2026-04-08	2024-01-03 to 2026-04-10
Backtest sessions	567	556

D.2 COMPARED CONFIGURATIONS AND BASELINES

Ablations and backbone. The five component-ablation rows of Table 2 keep the multi-view scoring interface and the final synthesis of Section 3.4. Without C3, MML context and the typed-event graph are kept, but the graph evidence enters as a direct correction without reliability weighting or residualization. Without C2 and C3, MML text, including graph-derived context, reaches the Sentiment Agent and no numerical correction is applied. Without C1, C2, and C3, retrieved MML context is also removed, while company narratives and multi-view scoring remain. The reliability-only and residualization-only rows each remove one of the two correction operations from the complete system and keep the other. The backbone row replaces DeepSeek-V4-Flash with Qwen3-Max in every LLM role, with the same prompts, data, graph construction, correction, and evaluation panel. The five component ablations and the backbone comparison belong to the frozen configuration of Appendix D.1.

Single-agent stock-analysis comparison. Both the separate single-agent row and the reported complete multi-agent row rely on DeepSeek-V4-Flash. The single-agent row uses one generalist stock-analysis agent to produce sentiment, technical, and fundamental scores and confidences, while retaining frozen Macro-Agent outputs and typed-event correction. Table 2 reports the single-agent run's stored final score and the complete multi-agent system's score on their common U.S. panel.

Baselines. The 12–1 momentum score (Jegadeesh & Titman, 1993) is $\widetilde{O}_{i,t-21}/\widetilde{O}_{i,t-252} - 1$ on adjusted opens, available where 252 sessions of prices exist. MEME, R&D-Agent-Quant, and AI Hedge Fund are evaluated on the same data, period, return labels, and dated universe. All four baselines use the validity rules, portfolio rule, and trading costs of RICE-ALPHA (Appendix D.3), so their strategy rows in Table 1 differ from RICE-ALPHA only in the score. The daily net asset values of MEME and R&D-Agent-Quant are rebased on the session before the first execution day and aligned to the study's execution sessions (Eq. 14).

D.3 METRICS AND STATISTICAL TESTS

Target and validity. With $\widetilde{O}_{i,t}$ the adjusted open of stock i on trading date t, the target is the five-session return from the next open,

$$y_{i,t} = \widetilde{O}_{i,t+6}/\widetilde{O}_{i,t+1} - 1, \tag{13}$$

counted in trading sessions. It becomes observable at the open of $t + 6$, and historical feedback uses it only after that open precedes the scoring cutoff. Missing entry or exit prices make the target unavailable rather than zero. A daily cross section is valid for a method if it has at least 20 stocks with finite, nonconstant scores and targets; valid observations may therefore differ across methods.

Portfolio construction. All strategies in Table 1, including every baseline, use one long-only rule. On each weekly rebalancing date t, the portfolio of method m is reset to equal weights on the top 10% of constituents ranked by $a^m_{i,t}$, trades at the open of $t + 1$, and holds these stocks until the next rebalance, so the weekly holding period aligns with the five-session target of Eq. 13. The portfolio is fully invested and uses no leverage, short sales, stop-losses, or hedging. Each side of every trade pays 4+4 bp in the U.S. and 10+4 bp in Hong Kong (commission plus slippage); Appendix E.4 reports turnover and cost sensitivity.

Signal and portfolio metrics. For method m with score $a_{i,t}^m$ ($a_{i,t}^{\mathrm{RICE}} = a_{i,t}$ from Eq. 4) on its valid cross section $\mathcal{U}_t^m$, $\mathrm{IC}_t^m = \mathrm{Corr}_{i\in\mathcal{U}_t^m}(a_{i,t}^m, y_{i,t})$, and RankIC_t^m is the same correlation on average ranks. IC and RankIC are daily means over valid dates; ICIR and RankICIR divide them by the sample standard deviation and are not annualized. Undefined correlations are omitted, not set to zero. With net daily return R_t^m and $V_t^m = \prod_{s\le t}(1+R_s^m)$, $V_0^m = 1$, over T sessions,

$$\begin{aligned} \mathrm{ARR}^m &= (V_T^m)^{252/T} - 1, & \mathrm{SR}^m &= \sqrt{252}\,\overline{R}^m \big/ \mathrm{sd}_{T-1}(R_t^m), \\ \mathrm{MDD}^m &= -\min_t \big[V_t^m \big/ \max_{0\le s\le t} V_s^m - 1\big], & \mathrm{CR}^m &= \mathrm{ARR}^m \big/ \mathrm{MDD}^m, \end{aligned} \tag{14}$$

with a zero risk-free rate, net returns after the study's trading costs, MDD reported as a positive loss, and CR undefined when MDD is zero. The NDX and HSI references start at unit value at the first execution-day open and use open-to-open price returns over the same sessions, without costs or dividend reinvestment.

Statistical tests. Daily IC and RankIC overlap through the five-session target, so their t-statistics use Newey–West standard errors (Newey & West, 1987) with five lags (twenty in the sensitivity check). The component and baseline paired contrasts subtract a comparator's daily correlation from that of RICE-ALPHA on common valid dates and test the mean difference with a one-sided paired block bootstrap (Künsch, 1989) (20-session blocks, 5,000 replications, centered at the observed mean). Holm adjustment (Holm, 1979) is applied within each market and metric; the U.S. family has the five component-ablation rows and AI Hedge Fund, and the Hong Kong family has AI Hedge Fund. The separate complete multi-agent versus single-agent comparison uses the same 20-session circular blocks and 5,000 replications. The pooled summary averages the two markets equally on dates valid in both and uses five Newey–West lags over this common-date series.

E SUPPLEMENTARY ANALYSES

These analyses reuse the predictions and evaluation panels of Tables 1–3.

E.1 SIGNIFICANCE OF THE DAILY CORRELATION SERIES

Table 9 reports Newey–West t-statistics for the RICE-ALPHA daily IC and RankIC series and the paired increments over the five U.S. ablation rows and AI Hedge Fund (Appendix D.3). With 20 lags, the IC t-statistics remain 4.72 in the U.S. and 2.28 in Hong Kong. The increments over the five U.S. ablation rows and over AI Hedge Fund in both markets are significant after Holm adjustment. Every compared configuration was frozen before the evaluation period (Appendix D.1), so the tests involve no selection on 2024–2026 data.

E.2 UNIVERSE ROBUSTNESS

Table 10 removes a fixed group of ten constituents from each daily cross section: AAPL, MSFT, NVDA, AMZN, GOOGL, GOOG, META, TSLA, AVGO, and COST in the U.S., and 00700.HK, 09988.HK, 00005.HK, 01299.HK, 00939.HK, 03690.HK, 00941.HK, 01398.HK, 02318.HK, and 00388.HK in Hong Kong. IC and RankIC remain positive and significant in both markets.

E.3 QUINTILE RETURNS AND HORIZON PROFILE

In Figure 6(a), the top-minus-bottom quintile spread in mean five-session forward return is 45 bp in the U.S. ($t = 3.59$) and 26 bp in Hong Kong ($t = 1.39$). In Figure 6(b), U.S. IC is 0.0159, 0.0413, and 0.0396 at one, ten, and twenty sessions, and Hong Kong IC is 0.0164, 0.0292, and 0.0310, so IC at ten and twenty sessions exceeds the one-session IC in both markets. The U.S. ablation row without C3 has lower IC than RICE-ALPHA at every evaluated horizon, and in the U.S. RICE-ALPHA also exceeds momentum at five, ten, and twenty sessions.

E.4 TRADING-COST SENSITIVITY

Annual one-way turnover under the study setting is 8.2 times NAV in the U.S. and 4.4 times NAV in Hong Kong. In Table 11, the U.S. Sharpe ratio is 1.551 with ARR 28.60% at 20 bp per side, and the

Table 9: Daily signal significance and mechanism contrasts. Panel A gives mean IC and RankIC with Newey–West t-statistics (lag 5; 20-lag IC values 4.72 and 2.28). Pooled means weight the two markets equally on 534 shared valid dates. Panel B compares the complete system with six U.S. rows or Hong Kong AI Hedge Fund on common valid dates; p-values use one-sided 20-session paired block bootstrap (5,000 replications), with Holm adjustment by market and metric. Panel C reports the separate unadjusted comparison of reliability-only against residualization-only; it is outside the Panel B Holm family.

Panel A: RICE-ALPHA *means*

Market	Dates	IC (t)	RankIC (t)
U.S.	562	0.0350 (4.31)	0.0360 (4.24)
Hong Kong	551	0.0283 (2.48)	0.0395 (3.10)
Pooled	534	0.0317 (4.17)	0.0377 (4.41)

Panel B: paired increments of RICE-ALPHA *over each comparator*

Market	Comparator	Metric	Increment	t_{diff}	p_{boot} (p_{Holm})
U.S.	w/o C1–C2–C3	IC	0.0286	3.80	0.0002 (0.0012)
		RankIC	0.0270	3.40	0.0004 (0.0024)
	w/o C2–C3 (memory only)	IC	0.0159	3.70	0.0002 (0.0012)
		RankIC	0.0155	3.34	0.0004 (0.0024)
	w/o C3 (direct correction)	IC	0.0090	2.44	0.0096 (0.0192)
		RankIC	0.0070	1.82	0.0444 (0.0480)
	AI Hedge Fund	IC	0.0252	2.58	0.0024 (0.0072)
		RankIC	0.0246	2.64	0.0024 (0.0096)
	Residualization only	IC	0.0142	2.86	0.0008 (0.0032)
		RankIC	0.0118	2.27	0.0120 (0.0360)
	Reliability weighting only	IC	0.0023	1.66	0.0378 (0.0378)
		RankIC	0.0030	2.02	0.0240 (0.0480)
Hong Kong	AI Hedge Fund	IC	0.0305	2.52	0.0020 (0.0020)
		RankIC	0.0370	3.06	0.0004 (0.0004)

Panel C: reliability-only minus residualization-only (unadjusted)

Market	Contrast	Metric	Increment	t_{diff}	p_{boot}
U.S.	Reliability only − residualization only	IC	0.0119	2.35	0.0050
		RankIC	0.0087	1.67	0.0406

Table 10: Constituent-exclusion robustness for the RICE-ALPHA daily correlations. The fixed group of ten constituents removed from each market is listed in Appendix E.2. Entries are means with Newey–West t-statistics (lag 5) in parentheses.

Market	Sample	Dates	IC (t)	RankIC (t)
U.S.	RICE-ALPHA, full evaluation panel	562	0.0350 (4.31)	0.0360 (4.24)
	RICE-ALPHA, fixed ten-stock exclusion	562	0.0334 (4.08)	0.0355 (4.12)
Hong Kong	RICE-ALPHA, full evaluation panel	551	0.0283 (2.48)	0.0395 (3.10)
	RICE-ALPHA, fixed ten-stock exclusion	551	0.0316 (2.69)	0.0434 (3.29)

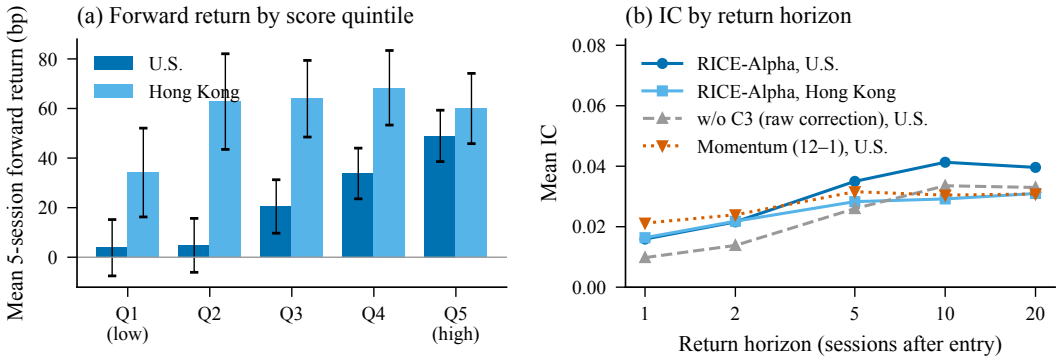


Figure 6: Signal diagnostics for the RICE-ALPHA score. (a) Mean five-session forward return by score quintile with standard errors over signal dates. (b) Mean IC against returns over one to twenty sessions after entry. The w/o C3 (raw correction) curve is the w/o C3 (direct) row of Table 2.

Hong Kong Sharpe ratio is 1.659 with ARR 35.03% at 28 bp per side; both exceed the corresponding baseline values under the study settings in Table 1.

Table 11: Sensitivity of the RICE-ALPHA portfolios to per-side trading costs. Cost per side includes commission and slippage; the study's setting (U.S. 4+4 bp, Hong Kong 10+4 bp) is marked.

Market	Cost per side (bp)	ARR (%)	Sharpe	MDD (%)	CR
U.S.	0	32.38	1.726	11.97	2.706
	4	31.62	1.691	12.03	2.629
	8 (study)	30.86	1.656	12.09	2.552
	12	30.11	1.621	12.16	2.476
	16	29.35	1.586	12.23	2.400
	20	28.60	1.551	12.30	2.325
Hong Kong	0	38.27	1.791	11.81	3.240
	7	37.45	1.758	11.82	3.168
	14 (study)	36.64	1.725	11.83	3.097
	21	35.83	1.692	11.84	3.027
	28	35.03	1.659	11.85	2.957

E.5 MARKET-CONDITION MEASURES

Table 3 and Section 4.4 use the daily net returns behind Table 1 and the index series of Appendix D.3. A session is an index-down (index-up) session when the index's open-to-open return is negative (positive); the first execution session, whose index return is zero by construction, is excluded. Capture is a strategy's mean return over these sessions divided by the index's, and beta is the least-squares slope of its daily net return on the index return. The crash window runs from the session after the index's peak to its trough, the sessions that define its maximum drawdown, and down months are calendar months with a negative compounded index return. Table 12 reports each strategy's market exposure and tail sessions. RICE-ALPHA beat the index in 6 of the 7 U.S. and all 13 Hong Kong index-down months.

Table 12: Market exposure and tail sessions. Beta: slope of daily net return on the index return. Worst/best decile: mean return (bp) on the index's worst and best tenth of sessions. Ratio: up-session capture divided by down-session capture.

	U.S. (Nasdaq-100)				Hong Kong (HSI)			
Method	Beta	Worst dec.	Best dec.	Ratio	Beta	Worst dec.	Best dec.	Ratio
Index (NDX / HSI)	1.00	−249	228	1.00	1.00	−247	285	1.00
MEME	0.20	−74	77	1.19	0.03	0	23	4.23
R&D-Agent-Quant	0.18	−74	75	1.22	0.28	−69	83	1.32
AI Hedge Fund	0.73	−174	163	0.96	0.96	−237	267	1.04
Momentum (12–1)	0.80	−205	168	1.02	0.53	−125	162	1.11
RICE-ALPHA (ours)	0.53	−128	124	1.27	0.46	−123	102	1.48

F LIMITATIONS

The study covers two large-cap equity universes. Evaluation-period selection can make estimates optimistic; tests of the selected series do not adjust for this selection. Event coverage and conditions may change results elsewhere.

Portfolio evaluation uses DeepSeek-V4-Flash; Qwen3-Max is evaluated only for prediction quality. Both models may contain information that postdates individual forecasts.

G AI USE STATEMENT

Generative AI tools assisted literature search and screening, manuscript organization, method–claim consistency review, language editing, and the implementation of experimental code. We take responsibility for the final manuscript, experimental implementation, and reported results, including content produced with AI assistance.