PARAFUSIONBENCH: A UNIFIED BENCHMARK FOR CROSS-PARADIGM FUSION IN FACTOR MINING AND TRADING DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantitative investment research has evolved from conventional models to large language models (LLMs) and multi-agent collaboration, forming two primary research trajectories: factor mining and trading decision-making. Their intersection yields four paradigms: Conventional Factor Mining (CFM), LLM-driven Factor Mining (LFM), Conventional Trading Decision (CTD), and LLM-driven Trading Decision (LTD). Existing works mostly develop individual paradigms in isolation, lacking unified formal characterization and cross-paradigm fusion evaluation. Current benchmarks similarly tend to focus on a single paradigm or task, limiting themselves to model ranking while offering little insight into paradigm fusion. To address this gap, we propose **ParaFusionBench**, a cross-paradigm fusion benchmark, and construct a quantitative paradigm fusion framework, including formal fusion definitions, a seven-dimensional feasibility criterion, fusion injection methods, candidate fusion screening methods, and fusion effect evaluation metrics. It designs the EDD three-layer evaluation system to guide multi-dimensional evaluation of different paradigm instances. Experimental validation shows that the complementary strengths and weaknesses of individual paradigms provide feasible conditions for paradigm fusion. 83% of fusion combinations help reduce drawdown, and 66.7% and 58.3% of combinations achieve Sharpe and ARR gains, respectively. ParaFusionBench also identifies effective combinations such as AlphaGen + TradingAgents and AlphaProbe + FinMem, demonstrating its ability to uncover beneficial cross-paradigm compositions. These results further provide practical insights into when and how heterogeneous quantitative paradigms can be effectively integrated. Our code is available at `https://anonymous.4open.science/r/ParaFusionBench-301F`.

1 INTRODUCTION

Quantitative investment aims to extract predictive signals from market data and construct trading strategies through systematic methods to obtain stable excess returns. In recent years, this field has undergone a profound evolution from traditional parametric models to large-scale pre-trained language models (Jiang et al., 2017; Chen et al., 2021; Fu, 2025). Traditional methods, represented by genetic programming, reinforcement learning, and deep learning, have achieved remarkable results in factor mining and trading decisions (Zhang et al., 2020; Cui et al., 2021; Yu et al., 2023; Zhu & Zhu, 2025). The rise of large language models (LLMs) has sparked a new paradigm in quantitative research, demonstrating unique advantages in factor discovery, logical verification, and trading decisions with powerful reasoning and generation capabilities (Wang et al., 2025; Li et al., 2024; Tang et al., 2025; Han et al., 2026). The introduction of agent technology further promotes quantitative research toward interpretability, collaboration, and dynamic adaptation (Yu et al., 2024; Xiao et al., 2025; Li et al., 2025c; Liu & Dang, 2025; Almon et al., 2025; Li et al., 2025b; Kou et al., 2025).

Despite these advances, existing quantitative research still lacks a unified, systematic, and insightful evaluation framework. Along the task dimension, methods target either factor mining or trading decisions; along the technical route dimension, methods use either conventional algorithms or LLMs. Their intersection naturally forms four research paradigms: Conventional Factor Mining (CFM), LLM-driven Factor Mining (LFM), Conventional Trading Decision (CTD), and LLM-driven Trading

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.

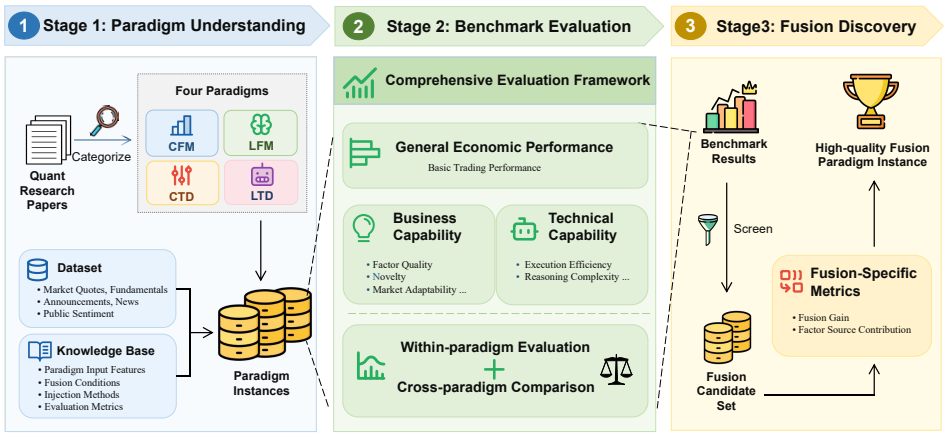


Figure 1: Framework of ParaFusionBench.

Decision (LTD). However, most existing works develop a single paradigm in isolation, lacking unified formal characterization and cross-paradigm comparison, and ignoring systematic exploration of inter-paradigm fusion and transfer effects.

In terms of evaluation benchmarks, several representative works exist. AlphaBench first systematically evaluates LLMs in formulaic alpha mining (Luo et al., 2026); AlphaEval proposes a five-dimensional factor evaluation framework (Ding et al., 2026); AlphaForgeBench defines LLMs as quantitative researchers, requiring executable factors and strategy construction (Zhang et al., 2026); InvestorBench focuses on evaluating LLM agents in diverse financial decision-making scenarios (Li et al., 2025a). However, these benchmarks mostly evaluate a single method or task. They fail to deeply investigate the internal mechanisms of each paradigm, lack capabilities for discovering paradigm fusion strategies, and cannot resolve the fusion matching problem between different factor mining and trading decision paradigms.

Therefore, to address this gap, we propose ParaFusionBench, a cross-paradigm fusion evaluation benchmark. It aims to evaluate existing quantitative research paradigms, screen out fusable paradigm combinations, and discover fused paradigms with gain through fusion effect evaluation (Figure 1). The main contributions are as follows:

(1) We propose a basic framework for quantitative paradigm fusion, including formal fusion definitions, a seven-dimensional fusion feasibility criterion, fusion injection methods, candidate fusion screening methods, and fusion effect evaluation metrics.

(2) We propose the cross-paradigm evaluation system EDD (Evaluation–Diagnosis–Discovery), covering a common performance layer, an intrinsic capability layer, and a fusion evaluation layer, to guide multi-dimensional evaluation of different paradigm instances.

(3) We conduct a systematic experimental evaluation on ParaFusionBench. Using 11 influential paradigm instances as objects, we validate the effectiveness of ParaFusionBench in paradigm fusion evaluation. 83% of fusion combinations help reduce drawdown, and 66.7% and 58.3% of combinations achieve Sharpe and ARR gains, respectively. Experiments reveal combinations with fusion gains such as AlphaGen+TradingAgents and AlphaProbe+FinMem.

(4) Important insights are obtained through the fusion process: fusion compatibility is more important than factor quality, and the influence of the downstream strategy side is greater than that of the upstream factor source; measures such as choosing the rule-guided injection method, moderately increasing the number of factors, and adopting the mixed type optimally will help improve the robustness and returns of fusion.

2 QUANTITATIVE TRADING PARADIGM FUSION FRAMEWORK

2.1 PARADIGM DEFINITION AND PARADIGM INSTANCES

General Formal Definition of Paradigms. In this paper, we define P as a research paradigm, which can be formally defined as the following quadruple:

$$P = \{I, O, L, \mathcal{E}\}$$

Where I: input space, denoting the data types that the paradigm can receive; O: output space, denoting the result types the paradigm produces; L: learning or optimization objective, denoting the objective function optimized during paradigm training or inference; $\mathcal{E}$: evaluation system, denoting the core metrics for measuring paradigm performance.

This abstract definition is not confined to a particular domain. Different research fields can assign specific meanings to I, O, L, and $\mathcal{E}$ according to their own characteristics.

A paradigm can have multiple concrete implementations, called **paradigm instances**. A paradigm instance is a concrete method, model, or system embodying a paradigm. Instances of the same paradigm share its abstract definition but may differ in input/output forms, optimization objectives, and evaluation metrics. This paper defines fusion at the paradigm level. Paradigm instance fusion applies paradigm fusion to concrete instances, following the same fusion forms, qualification conditions, and evaluation framework; it is not defined separately.

Four Quantitative Trading Paradigms. Based on task objective (factor mining vs. trading decision-making) and technical route (conventional models vs. LLMs), this paper divides quantitative trading research into four paradigms: CFM, LFM, CTD, and LTD. Following the quadruple formal definition above, their formal definitions can be readily specified; specific values for input, output, optimization objective, and evaluation system are given in Appendix A.

2.2 PARADIGM FUSION

General Definition of Paradigm Fusion. In this paper, paradigm fusion means that the output of one paradigm (the upstream paradigm) is used, after bridge mapping, as the input of another paradigm (the downstream paradigm), forming a new paradigm through fusion. Let the upstream paradigm be P_i and the downstream paradigm be P_j. Paradigm fusion is defined as: $P_{i\to j} = P_i \xrightarrow{B} P_j$. Here $\xrightarrow{B}$ denotes the fusion operator with bridge mapping, and B is the bridge mapping. The fused paradigm P_f can still be represented as a quadruple: $P_f = \{I_f, O_f, L_f, \mathcal{E}_f\}$ where: $I_f = I_i, O_f = O_j, L_f = \alpha L_i + \beta L_j + \gamma L_{\text{joint}}$.

The key to paradigm fusion lies in the existence of a bridging function B, which transforms the output of the upstream paradigm into the input of the downstream paradigm $B : O_i \to I_j$. The introduction of the bridge function enables the fused paradigm to handle inconsistencies between the upstream output and the downstream input, providing a formal basis for designing different injection forms.

Paradigm Fusion in Quantitative Trading. The four quantitative trading paradigms could form six theoretical fusion forms: CFM→CTD (conventional factors consumed by conventional decisions), CFM→LTD (conventional factors consumed by LLM decisions), LFM→CTD (LLM factors consumed by conventional decisions), LFM→LTD (LLM factors consumed by LLM decisions), CTD→CFM (trading decisions guide factor mining), and LTD→LFM (LLM decisions guide LLM factor mining).

2.3 PARADIGM FUSION FEASIBILITY CONDITIONS

General Fusion Criteria (FUSE-7). If two paradigms P_i and P_j can be fused, the following seven conditions must be satisfied: $Fuse(P_i, P_j) \Leftrightarrow I \wedge T \wedge G \wedge D \wedge E \wedge C \wedge R$, where: I denotes interface bridgeability; T denotes no temporal leakage; G denotes jointly optimizable objectives; D denotes alignable granularity; E denotes unified evaluation; C denotes causal consistency; R denotes feasible resources. This criterion is called FUSE-7 (Fusion Feasibility Seven-Dimensional Criteria). The item-by-item definitions are given in Appendix B. These conditions are general and domain-independent. Below will concretize them into operable screening criteria in light of specific domain characteristics.

Feasibility Assessment of Quantitative Fusion Directions. According to the FUSE-7 criterion, we evaluate the six fusion directions from Section 2.2 using a three-level scoring system: 2 (easy), 1 (conditional), 0 (difficult). Total scores sum seven items, with a maximum of 14. The scores of the above six fusion directions are as follows: CFM→CTD, 14 points; CFM→LTD, 9 points; LFM→CTD, 10 points; LFM→LTD, 8 points; CTD→CFM, 5 points; LTD→LFM, 3 points. Detailed scoring and results are in Appendix C.

For FM→TD, factor mining output naturally serves as trading decision input, enabling direct interface bridging. Their time order aligns, making temporal leakage easy to prevent. Objectives can be coordinated through joint or staged optimization; frequency and asset domains can be unified; and evaluation can use the same backtesting framework. Thus CFM→CTD is most feasible, followed by LFM→CTD, while LFM→LTD is slightly less feasible due to higher resource costs. For TD→FM, trading decision outputs actions or weights, which are hard to use directly as factor mining input. Feedback signals are noisy and non-stationary, complicating joint optimization; causal consistency and resource feasibility are also challenging. This direction therefore has low feasibility and is not the focus. Subsequent discussion centers on FM→TD.

2.4 Factor Injection Methods and Selection

In FM→TD fusion, the output of factor mining must be transformed into the input of trading decisions through a bridge mapping. This paper concretizes this bridge mapping as factor injection methods. Depending on the trading decision paradigm, factor injection can be divided into three main forms: structured value injection (VAL), structured value with semantic explanations (EXP), and structured value with rule-based decision governor (GOV). Examples, advantages, and disadvantages of these three injection methods are in Appendix D. Different injection paradigms lead to different fusion effects. Therefore, when fusing paradigms, one should consider the properties and characteristics of both upstream and downstream paradigms and select an appropriate injection method.

2.5 Evaluation Metrics for Paradigm Fusion

To evaluate the effect of paradigm fusion, it is necessary to formulate a series of evaluation metrics, mainly including the performance evaluation metrics of the paradigms themselves before fusion and the effect evaluation metrics of the fused paradigms. To this end, this paper proposes and formulates a three-layer evaluation metric system, the details of which are presented in Section 3.

3 Paradigm Evaluation Metric System and Fusion Screening

3.1 Overview of the EDD Evaluation System

In the section, we propose the EDD system (Evaluation–Diagnosis–Discovery Fusion Evaluation System), a unified benchmark for evaluating multiple paradigms in economic performance, task capability, technical efficiency, and fusion effect. It has three layers: (1) Common Performance Layer, providing comparable basic metrics; (2) Intrinsic Capability Layer, evaluating specialized capabilities by business capability and technical route; (3) Fusion Evaluation Layer, guiding fusion instance screening and evaluating factor-mining-to-trading-decision fusion, as shown in Figure 2.

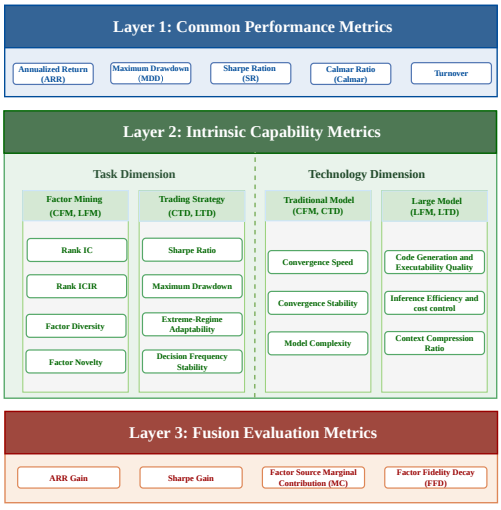


Figure 2: EDD metric evaluation system.

3.2 Metrics of the EDD Evaluation System

Common Performance Layer. The first layer provides a unified final acceptance benchmark,

answering the question "whether the paradigm
ultimately creates economic value," ensuring that all paradigms are fairly compared under the same economic standard. For factor mining paradigms (CFM, LFM), their output factor pools need to be transformed into backtestable top-k portfolios through a unified strategy wrapper (such as LightGBM-based aggregation), so as to ensure comparison with trading decision paradigms under the same framework.

Intrinsic Capability Layer. The second layer diagnoses why each paradigm performs well or poorly by evaluating task-specific capabilities in factor mining and trading, as well as the quality, efficiency, and cost of conventional and LLM-based methods. It supports cross-paradigm comparison and provides the basis for fusion analysis.

Fusion Evaluation Layer. The third layer evaluates inter-paradigm fusion and transfer effects, identifying which paradigm combinations yield performance gains and providing methodological guidance for cross-paradigm integration in quantitative research.

The indicators included in each layer of EDD and their calculation formulas are shown in Appendix E.

3.3 Screening Conditions for Paradigm Fusion Instances

To reduce the high cost of fusion experiments across numerous instance combinations, we propose TriSift, a three-stage screening and balancing method. It first applies hard thresholds to eliminate slim-chance combinations, then uses the first two layers of the EDD system (Common Performance Layer and Multi-Dimensional Capability Layer) for soft-condition ranking, and finally adjusts for coverage balance to select representative, reproducible, and balanced fusion combinations.

Hard-Threshold Filtering. Candidate combinations must satisfy all five Boolean conditions: (1) Direction matching: fusion direction must be FM→TD; (2) Interface bridgeable: factor output can be converted into decision input (numerical vectors, structured data, explanations, or tool calls); (3) Market and frequency consistency: same market and frequency (e.g., daily) with alignable time intervals; (4) No look-ahead bias: only historical information used; (5) Open-source reproducibility: code or reproducible protocol provided. Any failure leads to direct elimination.

Soft-Condition Screening (TriSift). Candidates are ranked using the first two EDD layers.

Step 1: Internal Comprehensive Ranking. Internal Comprehensive Ranking. Factor mining paradigms (CFM, LFM) and trading decision paradigms (CTD, LTD) are ranked separately using L1 and L2 metrics. For each metric, instances are sorted from best to worst: first place gets 1 point, second gets 2, etc. Reverse indicators (e.g., MDD, token cost, latency) are converted into positive efficiency rankings. Rankings across all metrics are averaged to obtain a comprehensive score; smaller is better.

Step 2: Fusion Combination Scoring. For each candidate, where F is the factor mining instance and T is the trading decision instance, define the screening score: $Score(F, T) = \alpha \cdot R_F + \beta \cdot R_T$ where: R_F is the internal composite ranking score of factor mining instance F; R_T is the internal composite ranking score of trading decision instance T. The weights are each set to 1/2 to balance the factor side and the decision side. A smaller score is better. In each cell (CFM/LFM × CTD/LTD), the lowest-scoring combination is preferred.

Step 3: Coverage Balance Adjustment. Coverage Balance Adjustment. Check coverage in the 2×2 matrix and ensure each cell has at least one representative, avoiding missing paradigm directions due to generally low scores. This method relies entirely on objective evaluations from the first two EDD layers. It considers investment performance, business capability, technical efficiency and cost, and fusion compatibility, ensuring selected combinations have both good single-paradigm performance and fusion potential. Through ranking and threshold screening, it narrows candidates under limited computation, reducing experiment costs. Specific criteria are in Appendix F.

4 EXPERIMENTAL EVALUATION AND RESULT ANALYSIS

4.1 EXPERIMENTAL OBJECTIVES AND SETUP

Experimental Objectives: We address three questions: RQ1—how each paradigm instance performs; RQ2—which instances to select for fusion and how fusion performs; RQ3—how injection method, factor quantity, and factor type affect fusion.

Experimental Setup: We evaluate 11 influential, open-source paradigm instances (AlphaGen, AlphaQCM, AlphaAgent, R&D-Agent, AlphaProbe, QuantaAlpha, FactorMiner, FinRL-PPO, FinRL-A2C, FinMem, TradingAgents) from ICLR, ICML, NeurIPS, KDD, or GitHub (>100k collections). Data comes from the U.S. market. Factor mining uses S&P 500 data; trading decisions use five widely used stocks—AAPL, GOOGL, MSFT, NVDA, TSLA (Yu et al., 2024; Xiao et al., 2025; Zhang et al., 2026). Sentiment and announcement data are from Yahoo Finance, Alpaca/Finnhub, and SEC EDGAR. A standard LightGBM + Top-K Dropout executor serves as the proxy task for factor-pool trading. Other environment details are in Appendix G.

4.2 RESULT ANALYSIS

4.2.1 RQ1: PERFORMANCE OF EACH PARADIGM INSTANCE ITSELF

Through experimental tests of each paradigm instance in terms of general investment performance, business tasks, and the application efficiency of technical models (as shown in Tables 1–4), the analysis reveals: (1) **Each paradigm instance shows high heterogeneity.** In LFM, AlphaProbe (ARR 33.64%, Sharpe 0.989) and AlphaAgent (32.83%, 0.936) lead, while FactorMiner loses (-20.84%). In LTD, TradingAgents excels (67.94%, 1.541), while FinMem is mediocre (12.29%, 0.497). CFM loses overall, whereas CTD is robust. Factor quality and performance decouple: AlphaGen has the highest Rank IC (0.0265) but loses; AlphaProbe (Rank IC 0.0107, novelty 0.12) performs best. (2) **Strengths and weaknesses of each paradigm instance in specific domains:** CFM has high predictive power but poor generalization. LFM generates novel factors but is costly (AlphaAgent Token Efficiency (TE) 133,390), and high entropy does not guarantee quality (QuantaAlpha entropy 0.9166 but negative Rank IC). CTD captures bull markets (FinRL-PPO Sharpe 2.908) but is weak in bear markets (-0.501). LTD adapts strongly (TradingAgents bear Sharpe 6.865; FinMem sideways 5.403) but is costly (TE 50,201) and slow (Sharpe per Hour 0.147).(3) **The complementarity of strengths and weaknesses provides feasible conditions for paradigm fusion.** Factor predictive power and strategy execution are complementary: factors from CFM/LFM can be injected into CTD/LTD executors. Technical efficiency complementarity: FactorMiner efficiently generates candidate factors (TE 971), AlphaAgent selects high-quality factors, and FinMem makes efficient decisions (TE 3,589).

Table 1: Overall economic performance of different paradigm approaches.

Paradigm	Instances	ARR%	| MDD% |	Sharpe	Calmar	Sortino
baseline	buy and hold	33.94	30.51	1.074	1.112	1.587
F. Pools	alpha20	6.59	38.51	0.273	0.287	0.548
	alpha158	-4.96	33.02	0.041	-0.002	0.092
CFM	AlphaGen	-23.59	40.17	-0.591	-0.583	-0.870
	AlphaQCM	-26.64	44.97	-0.685	-0.545	-0.791
LFM	RD-Agent	12.33	31.00	0.494	0.419	0.779
	QuantaAlpha	27.41	25.54	0.730	1.452	1.334
	AlphaAgent	32.83	32.26	0.936	1.427	1.354
	AlphaProbe	33.64	22.08	0.989	1.528	1.622
	FactorMiner	-20.84	44.68	-0.572	-0.397	-0.728
CTD	FinRL-A2C	26.22	29.14	0.919	0.939	1.426
	FinRL-PPO	27.59	20.84	1.098	1.390	1.673
LTD	TradingAgents	67.94	29.41	1.541	2.310	2.043
	FinMem	12.29	30.95	0.497	0.397	0.606

Table 2: Factor quality performance of the CFM and LFM paradigm instances.

Paradigm	Instances	Rank IC	Rank ICIR	Corr Entropy	Novel Ratio
F. Pools	alpha20	0.0036	0.030	—	—
	alpha158	0.0030	0.032	—	—
CFM	AlphaGen	0.0265	0.1720	0.7917	0.62
	AlphaQCM	0.0166	0.1528	0.5963	0.97
LFM	RD-Agent	0.0153	0.1004	0.6104	0.20
	QuantaAlpha	-0.0064	-0.0588	0.9166	0.55
	AlphaAgent	0.0167	0.1505	0.7246	0.60
	AlphaProbe	0.0107	0.0866	0.6997	0.12
	FactorMiner	0.0072	0.0607	0.9256	0.12

Table 3: Trading decision performance of the CTD and LTD paradigms. Bull/Bear/Side/Vol: Sharpe ratio (SR) in bull (n=121), bear (n=37), sideways (n=92), and high-volatility (n=125) regimes; SR/h: SR per hour.

Paradigm	Instances	Bull	Bear	Side	Vol	SR/h
baseline	buy and hold	1.765	1.268	0.524	0.967	—
F. Pools	alpha20	2.455	-1.243	-1.656	-0.384	—
	alpha158	0.053	-0.137	0.209	-0.878	—
CTD	FinRL-A2C	2.838	-1.141	0.727	0.322	—
	FinRL-PPO	2.908	-0.501	0.391	0.634	—
LTD	TradingAgents	3.251	6.865	0.355	1.128	0.147
	FinMem	-1.390	-1.910	5.403	0.131	0.055

Table 4: Execution efficiency of large model methods in factor mining and trading decisions. Succ.: execution success rate; Tokens: total token usage; In/Out: average input/output tokens; CCR: context compression ratio.

Paradigm	Instances	Succ.	Tokens	In	Out	CCR
LFM	RD-Agent	0.454	71,729	29,168	3,056	0.105
	QuantaAlpha	0.500	26,570	5,902	7,383	1.254
	AlphaAgent	0.861	133,390	105,431	8,826	0.084
	AlphaProbe	0.046	37,911	1,317	405	0.308
	FactorMiner	0.111	1,303	100	35	0.347
LTD	TradingAgents	1.000	50,201	12,275	6,561	0.535
	FinMem	1.000	3,589	2,778	446	0.160

4.2.2 RQ2: Analysis of Fusion Effects of Different Paradigm Instances

Based on the results of the previous experiments, using the TriSift method, four upstream paradigm fusion instances (AlphaGen, AlphaQCM, AlphaProbe, AlphaAgent) and three downstream paradigm fusion instances (FinRL-PPO, TradingAgents, FinMem) can be selected. They can form 12 possible paradigm fusion combinations. The screening process is in Appendix F and supplementary figures are provided in Appendix H. Table 5 presents the 12 fusion results:

(1) **Paradigms achieve better returns through appropriate fusion.** 83% (10/12) fusion combinations improve MDD, but only 66.7% (8/12) improve Sharpe and 58.3% (7/12) improve ARR. AlphaGen+TradingAgents in CFM+LTD performs best (ΔSharpe +1.116, ΔARR +87.13%), AlphaProbe+FinMem in LFM+LTD also performs excellently. However, we also observe that in some scenarios, certain fusions can also produce negative effects. For example, AlphaProbe+TradingAgents in LFM+LTD has a ΔSharpe of -3.193, and a similar example is AlphaAgent+TradingAgents. (2) **The downstream strategy side has a greater influence than the upstream factor source.** The same factor source fused with different downstream ends yields different results: FinMem achieves positive gains with all factor sources, showing the strongest compatibility; TradingAgents succeeds only with CFM factors and fails with LFM factors. Upstream Rank IC has no direct relationship with fusion success: AlphaGen has the highest Rank IC (0.0295) but mediocre fusion, while AlphaProbe has a Rank IC of only 0.0089 yet fuses successfully with FinRL-PPO and FinMem. Compatibility is more important than factor quality. (3) **Complementary feasibility cannot guarantee fusion success.** Under unified fusion parameters, adaptability differences lead to failure, e.g., TradingAgents is highly sensitive to prompt injection, and volatility-type factors conflict with it. Therefore EDD three-layer evaluation and TriSift screening are essential for systematically identifying highly compatible combinations and avoiding negative transfer.

4.2.3 RQ3: Analysis of Factors Influencing Paradigm Fusion

In the AlphaProbe+FinMem fusion ablation, injection method, factor quantity, and factor type significantly affect fusion gains. (1) In terms of injection method, **GOV yields the largest gains** (ΔSharpe +4.70, ΔARRpp +255.95), followed by EXP (ΔSharpe +3.94, ΔARRpp +231.74), with VAL being the lowest (ΔSharpe +3.70, ΔARRpp +208.96). Although EXP+GOV is on par with GOV in ΔMDDpp, its ΔSharpe and ΔARRpp are lower than those of GOV, indicating that decision constraints are key to improving fusion gain, and economic explanations also provide gains, but the combination of the two does not produce additional effects. (2) **Increasing factor quantity improves all gains but with diminishing returns:** K=1$\rightarrow$5 gives ΔSharpe +0.443 and ΔARRpp +27.87, while K=5$\rightarrow$10 only gives +0.049 and +11.67; This indicates that moderately increasing the number of factors helps improve the fusion effect, but too many factors may introduce noise, leading to diminishing marginal returns. (3) In terms of factor type, **the mixed type is optimal** (ΔSharpe +4.702, ΔARRpp +255.95), followed by the momentum type (ΔSharpe +4.175, ΔARRpp +207.80), with the price-volume type being the weakest (ΔSharpe +3.464, ΔARRpp +149.85). The mixed type leads across all metrics, indicating that the diversity of factor types helps provide complementary information and improves the robustness and returns of fusion.

Table 5: Comparison of fusion effects of different types across the four paradigms.

Paradigm	Instances	Rank IC	Sharpe	ARR%	MDD%	ΔSharpe	ΔARR%	ΔMDD	MC	FFD
baseline	alpha20	−0.0001	−0.346	−20.92	−44.66	–	–	–	–	–
CFM	AlphaGen	0.0295	−0.005	−1.92	−25.98	–	–	–	–	–
	AlphaQCM	0.0095	−1.876	−55.43	−38.58	–	–	–	–	–
LFM	AlphaAgent	0.0169	−0.079	−4.67	−25.35	–	–	–	–	–
	AlphaProbe	0.0089	2.350	153.03	−11.20	–	–	–	–	–
CTD	FinRL-PPO	–	0.456	10.23	−20.16	–	–	–	–	–
LTD	TradingAgents	–	0.653	22.08	−29.41	–	–	–	–	–
	FinMem	–	−0.041	−5.91	−21.13	–	–	–	–	–
CFM+CTD	AlphaGen+FinRL-PPO	−0.0311	0.386	10.40	−13.04	−0.069	+0.17	+7.12	−0.20	–
	AlphaQCM+FinRL-PPO	−0.0198	0.076	−0.81	−16.84	−0.380	−11.04	+3.32	+0.25	–
CFM+LTD	AlphaGen+TradingAgents	−0.0311	1.769	109.21	−19.59	+1.116	+87.13	+9.82	3.27	0.238
	AlphaQCM+TradingAgents	−0.0198	1.029	43.03	−28.06	+0.376	+20.95	+1.35	−0.25	0.728
	AlphaGen+FinMem	−0.0311	0.484	10.26	−16.53	+0.526	+16.17	+4.60	1.54	0.080
	AlphaQCM+FinMem	−0.0198	0.811	15.85	−14.81	+0.852	+21.76	+6.31	−0.56	0.605
LFM+CTD	AlphaAgent+FinRL-PPO	0.0191	0.850	8.34	−10.49	+0.394	−1.89	+9.67	1.48	–
	AlphaProbe+FinRL-PPO	−0.0508	0.666	7.58	−10.76	+0.211	−2.65	+9.40	0.08	–
LFM+LTD	AlphaAgent+TradingAgents	0.0191	−2.108	−62.71	−40.30	−2.761	−84.79	−10.89	−10.34	0.394
	AlphaProbe+TradingAgents	−0.0508	−2.541	−66.66	−46.95	−3.193	−88.74	−17.55	−1.18	0.382
	AlphaAgent+FinMem	0.0191	0.887	29.24	−15.70	+0.928	+35.15	+5.42	3.48	0.079
	AlphaProbe+FinMem	−0.0508	1.223	35.52	−15.94	+1.264	+41.43	+5.19	0.47	0.091

Table 6: Effects of different factors on paradigm fusion.

Dimension	Setting	Rank IC	Sharpe	ARR%	MDD%	ΔSharpe	ΔARR%	ΔMDD
baseline	no injection (native FinMem)	–	−0.532	−20.28	−18.28	–	–	–
Injection method	VAL (values only)	−0.0031	3.173	+188.69	−15.60	+3.70	+208.96	+2.67
	EXP (values + explanation)	−0.0031	3.412	+211.46	−14.65	+3.94	+231.74	+3.63
	GOV (values + constraints)	−0.0031	4.171	+235.67	−12.83	+4.70	+255.95	+5.44
	EXP+GOV (values + expl. + constr.)	−0.0031	3.664	+201.75	−12.83	+4.20	+222.03	+5.44
Factor count	K=1 cross-type	−0.0285	2.857	+114.68	−13.96	+3.39	+134.96	+4.32
	K=5 cross-type	−0.0212	3.300	+142.55	−13.96	+3.83	+162.83	+4.32
	K=10 cross-type	−0.0002	3.349	+154.22	−12.83	+3.88	+174.50	+5.44
Factor type	momentum (5)	−0.0031	3.644	+187.52	−13.69	+4.18	+207.80	+4.59
	volatility (5)	−0.0057	3.347	+189.61	−12.83	+3.88	+209.89	+5.44
	price-volume (5)	−0.0015	2.933	+129.57	−14.70	+3.46	+149.85	+3.57
	mixed (5, = GOV row)	−0.0031	4.171	+235.67	−12.83	+4.70	+255.95	+5.44

4.3 Analysis and Discussion

The complementary strengths and weaknesses of individual paradigms provide feasible conditions for paradigm fusion, and the experiments validate the value and effectiveness of ParaFusionBench.

Key insight: (1) Paradigms achieve better returns through appropriate fusion. For instance, AlphaGen+TradingAgents achieved a 1.116 Sharpe gain. (2) fusion compatibility is more important than factor quality. AlphaGen has the highest Rank IC (0.0295), yet its fusion performance is mediocre; AlphaProbe has a Rank IC of only 0.0089, yet it fuses successfully with FinRL-PPO and FinMem. (3) The influence of the downstream decision side is greater than that of the upstream factor source: FinMem achieves positive gains when fused with all factor sources, while TradingAgents succeeds only with CFM factors. (4) Measures such as choosing the GOV injection method, moderately increasing the number of factors, and adopting the mixed type optimally will help improve the robustness and returns of fusion.

Limitations: Current research on quantitative paradigm fusion is relatively scarce. Due to time and cost constraints, the fusion experiments only test on five stocks commonly used in research; the fusion direction is limited to FM→TD, while TD→FM has not yet been explored.

Future Research Suggestions: expand samples and market coverage; explore bidirectional closed-loop fusion; design dynamic fusion mechanisms that switch factor sources or decision sides according

to market regimes; optimize token efficiency and adaptive explanation mechanisms; and compare direct bridging with agent-mediated connection modes.

5 Related Work

Quantitative Trading Research. (1) **Conventional model-based factor mining.** Zhang et al. (2020) proposed AutoAlpha, Cui et al. (2021) proposed AlphaEvolve, which generate formulas through expression tree mutation; Yu et al. (2023) proposed AlphaGen, Zhu & Zhu (2025) proposed AlphaQCM, Tang et al. (2026) proposed AlphaAgentEvo, Shi et al. (2025a) proposed AlphaForge, and Zhao et al. (2025) proposed QFR, modeling factor generation as MDPs; Yang et al. (2026) proposed AlphaCFG, Chen et al. (2026) proposed AlphaSAGE, and Guo et al. (2026) proposed AlphaPROBE, introducing grammar, GFlowNet, and DAG. Traditional methods balance search efficiency and interpretability, but rely on predefined operators, have limited generalization, and are prone to overfitting and factor decay. (2) **LLM-based factor mining.** Wang et al. (2025) proposed AlphaGPT, Li et al. (2024) proposed FAMA, Tang et al. (2025) proposed AlphaAgent, Li et al. (2025d) proposed R&D-Agent(Q), Han et al. (2026) proposed QuantaAlpha, Liu et al. (2026) proposed CogAlpha, and Wang et al. (2026) proposed FactorMiner. They generate factors via prompt engineering, knowledge retrieval, experience memory, multi-agent collaboration, and evolutionary search (Chen & Kawashima, 2025; Shi et al., 2025b; Lin et al., 2026). They break through manual operator limits and improve diversity and anti-decay capability, but face hallucination, poor reproducibility, high evaluation cost, and prompt sensitivity. (3) **Traditional model-based trading decisions.** Liu et al. (2021) proposed FinRL and Liu et al. (2022) proposed FinRL-Meta; DeepTrader, SAC, PPO, DQN, and AlphaMix+ model trading as MDPs to learn buy/sell and portfolio weights. Qlib (Yang et al., 2020) and FinRL-Meta (Liu et al., 2022) provide data, backtesting, and training environments. They adapt to dynamic markets, but have low sample efficiency, weak interpretability, unstable under regime shifts, and rely on human experience for reward design and risk constraints. (4) **LLM-based trading decisions.** Yu et al. (2025) proposed FinMem, Zhang et al. (2024) proposed FinAgent, Yu et al. (2024) proposed FinCon, Xiao et al. (2025) proposed TradingAgents, Li et al. (2025c) proposed QuantAgents, Papadakis et al. (2025) proposed ATLAS, and Tian et al. (2025) proposed TradingGroup. They simulate investment teams, risk control, and dynamic prompt optimization. However, current agent systems coordinating multiple roles only provide fusion capability, but cannot answer "which combination of factor mining and trading decision methods is optimal," lacking meta-evaluation, causal attribution, and systematic comparison over the combination space.

Evaluation Benchmarks Research. Traditional evaluation relies on IC, RankIC, ICIR, and backtesting metrics. Yang et al. (2020) proposed Qlib, Liu et al. (2022) proposed FinRL-Meta, and Sun et al. (2023) proposed PRUDEX-Compass, providing data, environments, and multi-dimensional evaluation. For LLMs and agents, Luo et al. (2026) proposed AlphaBench, Ding et al. (2026) proposed AlphaEval, Zhang et al. (2026) proposed AlphaForgeBench, and Li et al. (2025a) proposed InvestorBench. AlphaBench systematically evaluates factor generation, evaluation, and search; AlphaEval assesses five dimensions: predictive power, stability, robustness, financial logic, and diversity; AlphaForgeBench emphasizes end-to-end strategy code generation and deterministic backtesting; InvestorBench covers single-asset tasks such as stocks, cryptocurrencies, and ETFs. However, existing benchmarks mostly evaluate a single method or task, do not deeply examine the internal mechanisms of each paradigm, lack insight discovery functions, and cannot systematically answer the fusion matching problem between different factor mining and trading decision methods.

6 Conclusion

This paper addresses the problems of paradigm fragmentation and the lack of fusion evaluation in AI quantitative research, and proposes ParaFusionBench, a cross-paradigm fusion evaluation benchmark. It constructs a paradigm fusion framework, the EDD three-layer evaluation system, and the TriSift screening method to discover paradigm fusions with gain benefits. Experiments validate the effectiveness of ParaFusionBench in paradigm fusion evaluation and discovery, yielding valuable fusion insights and future research suggestions, and advancing quantitative research from single paradigms to paradigm fusion.

AI USE STATEMENT

In this work, we used generative AI tools for interpreting results: the tools assisted the authors in summarizing and interpreting the outcomes of the benchmark experiments. The experiments themselves were designed, implemented, and run entirely by the authors, and all reported numbers come directly from these runs. We have not used generative AI tools for the other required-disclosure tasks, such as proposing or refining hypotheses, designing or providing feedback on the research methodology or experiments, implementing methods, cleaning or reformatting datasets, assisting with translation, supporting qualitative and thematic data analysis, or helping develop theoretical models, and generating synthetic datasets and formulating or proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools for recommended-disclosure tasks: editing the paper to improve readability (polishing the English wording) and assisting with formatting the manuscript (converting formula objects from the source manuscript into LaTeX and formatting tables and references). We have reviewed all AI-assisted work: every AI-assisted interpretation was checked against the benchmark runs, and every AI-assisted edit and transcribed formula was manually verified against the source files by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies quantitative investment methods through historical-data evaluation and does not involve human subjects or direct interaction with financial markets. The benchmark is intended solely for research purposes and should not be interpreted as financial or investment advice. Results obtained under historical backtesting may not generalize to live-market conditions, and deployment of the evaluated methods may involve financial risks. We report the data sources, evaluation assumptions, and experimental limitations to facilitate appropriate interpretation of the results.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce the main results of ParaFusionBench throughout the paper and supplementary materials. Section G.5 describes the experimental protocol and evaluated paradigm instances, while Appendix G reports the data configuration, implementation details, and experimental settings. The definitions of paradigm fusion, evaluation metrics, injection mechanisms, and TriSift screening criteria are provided in the corresponding methodological sections and appendices. We additionally provide the implementation code, experiment configurations, and supporting materials through the anonymized supplementary repository. Together, these materials are intended to enable reproduction of the standalone evaluation, fusion experiments, and ablation studies reported in this paper.

REFERENCES

Blake Almon, Matthew Caliboso, Alex Kim, Rohan Dutta, Rohan Raman, Mithil Srungarapu, Vasu Sharma, and Kevin Zhu. RAPTOR: Reasoned agentic portfolio trading with orchestrated rebalancing. In *NORA: The First Workshop on Knowledge Graphs & Agentic Systems Interplay*, 2025. URL https://openreview.net/forum?id=ziuTkKhgT0.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zRKF4ln2VE.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=a7APmM4B9d.

Qizhao Chen and Hiroaki Kawashima. Multi-agent llm framework for formulaic alpha generation and selection in quantitative trading. In *2025 IEEE International Conference on Big Data (BigData)*, pp. 7143–7152. IEEE, 2025.

Can Cui, Wei Wang, Meihui Zhang, Gang Chen, Zhaojing Luo, and Beng Chin Ooi. Alphaevolve: A learning framework to discover novel alphas in quantitative investment. In *Proceedings of the 2021 International conference on management of data*, pp. 2208–2216, 2021.

Hongjun Ding, Binqi Chen, Jinsheng Huang, Taian Guo, Zhengyang Mao, Guoyi Shao, Lutong Zou, Luchen Liu, and Ming Zhang. Alphaeval: A comprehensive and efficient evaluation framework for formula alpha mining. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 897–908, 2026.

Weilong Fu. The new quant: A survey of large language models in financial prediction and trading. *arXiv preprint arXiv:2510.05533*, 2025.

Taian Guo, Haiyang Shen, Junyu Luo, Binqi Chen, Hongjun Ding, Jinsheng Huang, Luchen Liu, Yun Ma, and Ming Zhang. Alphaprobe: Alpha mining via principled retrieval and on-graph biased evolution. *arXiv preprint arXiv:2602.11917*, 2026.

Jun Han, Shuo Zhang, Wei Li, Yifan Dong, Tu Hu, Yumo Zhu, Xiaomin Yu, Xin Guo, Zhaowei Liu, Kunyi Wang, Jingping Liu, Tianyi Jiang, Ruichuan An, Sen Hu, Zhi Yang, Ronghao Che, and Huacan Wang. Quantaalpha: An evolutionary framework for llm-driven alpha mining. *arXiv preprint arXiv:2602.07085*, 2026.

Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*, 2017.

Zhizhuo Kou, Holam Yu, Junyu Luo, Jingshu Peng, Xujia Li, Chengzhong Liu, Juntao Dai, Lei Chen, Sirui Han, and Yike Guo. Automate strategy finding with LLM in quant investment. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18517–18533, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1005. URL https://aclanthology.org/2025.findings-emnlp.1005/.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K.P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. INVESTORBENCH: A benchmark for financial decision-making tasks with LLM-based agent. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2509–2525, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.126. URL https://aclanthology.org/2025.acl-long.126/.

Jifeng Li, Arnav Grover, Abraham Alpuerto, Yupeng Cao, and Xiao-Yang Liu. Orchestration framework for financial agents: From algorithmic trading to agentic trading. *arXiv preprint arXiv:2512.02227*, 2025b.

Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. QuantAgents: Towards multi-agent financial system via simulated trading. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17438–17464, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.945. URL https://aclanthology.org/2025.findings-emnlp.945/.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025d. doi: 10.52202/085713-3967. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ac5c2b6e423883cbcacbcccf88491b78-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhiwei Li, Ran Song, Caihong Sun, Wei Xu, Zhengtao Yu, and Ji-Rong Wen. Can large language models mine interpretable financial factors more effectively? a neural-symbolic factor mining agent

model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3891–3902, 2024.

Qinhong Lin, Ruitao Feng, Yinglun Feng, Zhenxin Huang, Yukun Chen, Zhongliang Yang, Linna Zhou, Binjie Fei, Jiaqi Liu, and Yu Li. Factorengine: A program-level knowledge-infused factor mining framework for quantitative investment. *arXiv preprint arXiv:2603.16365*, 2026.

Bijia Liu and Ronghao Dang. Finrs: A risk-sensitive trading framework for real financial markets. *arXiv preprint arXiv:2511.12599*, 2025.

Fengyuan Liu, Yi Huang, Sichun Luo, Yuqi Wang, Yazheng Yang, Xinye Li, Zefa Hu, Junlan Feng, and Qi Liu. Cognitive alpha mining via llm-driven code-based evolution. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11715–11749, 2026.

Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and Christina Dan Wang. Finrl: Deep reinforcement learning framework to automate trading in quantitative finance. In *Proceedings of the second ACM international conference on AI in finance*, pp. 1–9, 2021.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 1835–1849. Curran Associates, Inc., 2022. doi: 10.52202/068431-0134. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0bf54b80686d2c4dc0808c2e98d430f7-Paper-Datasets_and_Benchmarks.pdf.

Haochen Luo, Ho Tin Ko, Jiandong Chen, David Sun, Yuan Zhang, and Chen Liu. Alphabench: Benchmarking large language models in formulaic alpha factor mining. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=d97Q8r7ZKZ.

Charidimos Papadakis, Angeliki Dimitriou, Giorgos Filandrianos, Maria Lymperaiou, Konstantinos Thomas, and Giorgos Stamou. Atlas: Adaptive trading with llm agents through dynamic prompt optimization and multi-agent coordination. *arXiv preprint arXiv:2510.15949*, 2025.

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, 2025a.

Yu Shi, Yitong Duan, and Jian Li. Navigating the alpha jungle: an llm-powered mcts framework for formulaic factor mining. *arXiv preprint arXiv:2505.11122*, 2025b.

Shuo Sun, Molei Qin, Xinrun Wang, and Bo An. Prudex-compass: Towards systematic evaluation of reinforcement learning in financial markets. *arXiv preprint arXiv:2302.00586*, 2023.

Ziyi Tang, Zechuan Chen, Jiarui Yang, Jiayao Mai, Yongsen Zheng, Keze Wang, Jinrui Chen, and Liang Lin. Alphaagent: Llm-driven alpha mining with regularized exploration to counteract alpha decay. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2813–2822, 2025.

Ziyi Tang, Xuexiong Yin, Weixing Chen, Zechuan Chen, Yongsen Zheng, Wenxuan Ye, Keze Wang, and Liang Lin. Alphaagentevo: Evolution-oriented alpha mining via self-evolving agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Feng Tian, Flora D Salim, and Hao Xue. Tradinggroup: A multi-agent trading system with self-reflection and data-synthesis. *arXiv preprint arXiv:2508.17565*, 2025.

Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Alpha-GPT: Human-AI interactive alpha mining for quantitative investment. In Ivan Habernal, Peter Schulam, and Jörg Tiedemann (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 196–206, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-334-0. doi: 10.18653/v1/2025.emnlp-demos.14. URL https://aclanthology.org/2025.emnlp-demos.14/.

Yanlong Wang, Jian Xu, Hongkang Zhang, Shao-Lun Huang, Danny Dongning Sun, and Xiao-Ping Zhang. Factorminer: A self-evolving agent with skills and experience memory for financial alpha discovery. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 12327–12338, 2026.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2025.

Han Yang, Dong Hao, Zhuohan Wang, Qi Shi, and Xingtong Li. Alpha discovery via grammar-guided learning and search. *arXiv preprint arXiv:2601.22119*, 2026.

Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. Qlib: An ai-oriented quantitative investment platform. *arXiv preprint arXiv:2009.11189*, 2020.

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.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 137010–137045. Curran Associates, Inc., 2024. doi: 10.52202/079017-4354. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f7ae4fe91d96f50abc2211f09b6a7e49-Paper-Conference.pdf.

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.

Tianping Zhang, Yuanqi Li, Yifei Jin, and Jian Li. Autoalpha: An efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment. *arXiv preprint arXiv:2002.08245*, 2020.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining*, pp. 4314–4325, 2024.

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.

Junjie Zhao, Chengxi Zhang, Min Qin, and Peng Yang. Quantfactor reinforce: mining steady formulaic alpha factors with variance-bounded reinforce. *IEEE Transactions on Signal Processing*, 73:2448–2463, 2025.

Zhoufan Zhu and Ke Zhu. Alphaqcm: Alpha discovery in finance with distributional reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025.

APPENDIX CONTENTS

A Formal Definitions of Four Quantitative Trading Paradigms 14

B Notation for the Criteria of Paradigm Fusion 15

C Analysis of the Seven-Dimensional Determination Conditions for Six Fusion Directions . . 16

D Factor Injection Methods and Their Characteristics 16

E Calculation Formulas for the Metrics of the EDD Evaluation System. 17

F Scoring and Screening of Paradigm Instance Fusion Combinations 20

G Experimental Setup Description 21

H Supplementary Experimental Results 23

A FORMAL DEFINITIONS OF FOUR QUANTITATIVE TRADING PARADIGMS

Based on the quadruple definition of a paradigm in Section 2.1, $P = (I, O, L, \mathcal{E})$, the specific values of the four quantitative trading paradigms are given in Table 7 and Table 8.

A.1 CONVENTIONAL FACTOR MINING (CFM) AND LLM-DRIVEN FACTOR MINING (LFM)

Table 7: Formal definitions of the CFM and LFM paradigms.

Paradigm Name	**Formal Expression**
Conventional Factor Mining (CFM)	• **Input space** I_{CFM}: Structured price-volume and financial data, such as opening price, closing price, high price, low price, trading volume, turnover rate, price-to-earnings ratio, etc. • **Output space** O_{CFM}: A set of factor expressions or a factor pool $F = \{f_1, f_2, \ldots, f_K\}$, where each factor f_k maps input data to a cross-sectional signal. • **Optimization objective** L_{CFM}: Factor predictive power, typically measured by the Information Coefficient (IC) or Rank Information Coefficient (RankIC): $L_{CFM} = E_i\left[IC(f_i, r_{i+1})\right]$ • **Evaluation system** $\mathcal{E}_{\mathrm{CFM}}$: IC, RankIC, ICIR, RankICIR, average pairwise factor correlation, factor coverage, etc.
LLM-driven Factor Mining (LFM)	• **Input space** I_{LFM}: Structured price–volume data and financial data, as well as unstructured textual data (news, research reports, financial report texts, etc.). • **Output space** O_{LFM}: A set of factor expressions or a factor pool F, usually accompanied by natural language explanations or semantic descriptions. • **Optimization objective** L_{LFM}: A joint objective of factor predictive power and semantic consistency: $L_{\mathrm{LFM}} = \alpha \cdot \mathrm{IC} + \beta \cdot \mathrm{Consistency}(h, d, f)$, where $\mathrm{Consistency}(h, d, f)$ measures the consistency among the market hypothesis, factor description, and factor expression. • **Evaluation system** $\mathcal{E}_{\mathrm{LFM}}$: IC, RankIC, ICIR, factor originality, expression complexity, semantic consistency score, token cost, inference latency, etc.

Table 8: Formal definitions of the CTD and LTD paradigms.

Paradigm Name	Formal Expression
Conventional Trading Decision (CTD)	• **Input space** I_{CTD}: Structured price–volume data, technical indicators, and factor value vectors. • **Output space** O_{CTD}: Trading actions (buy/sell/hold) or a portfolio weight vector $w \in \mathbb{R}^N$. • **Optimization objective** $L_{CTD} = E\left[\frac{\bar{r}-r_f}{\sigma(r)}\right]$: Risk-adjusted return, typically measured by the Sharpe Ratio or cumulative return. • **Evaluation system** $\mathcal{E}_{\mathrm{CTD}}$: Annualized return (ARR), Sharpe Ratio, maximum drawdown (MDD), Calmar Ratio, turnover, net return after costs, etc.
LLM-driven Trading Decision (LTD)	• **Input space** I_{LTD}: Structured price–volume data, technical indicators, factor value vectors, as well as unstructured textual data (news, announcements, social media sentiment, etc.). • **Output space** O_{LTD}: Trading actions or portfolio weights, usually accompanied by natural language decision rationales. • **Optimization objective** L_{LTD}: A joint objective of risk-adjusted return and decision interpretability: $L_{\mathrm{LTD}} = \alpha \cdot \mathrm{Sharpe} + \beta \cdot \mathrm{Explainability}$. • **Evaluation system** $\mathcal{E}_{\mathrm{LTD}}$: Annualized return, Sharpe Ratio, maximum drawdown, Calmar Ratio, decision stability, token cost, inference latency, etc.

A.2 Conventional Trading Decision (CTD) and LLM-driven Trading Decision (LTD)

B Notation for the Criteria of Paradigm Fusion

For the six fusion directions in Section 2.2, we analyze their satisfaction of the seven dimensions: interface bridgeable (I), temporal no-leakage (T), objective jointable (G), granularity alignable (D), evaluation unifiable (E), causal consistency (C), and resource feasible (R).

Table 9: Notation for the criteria of paradigm fusion.

Condition	Symbol	General Definition	Description
Bridgeable interface	I	$\exists B : O_i \to I_j$	The upstream output can be transformed into the downstream input via mapping.
Temporal no-leakage	T	$I_i(t) \subseteq \alpha\big[O_i(\leq t), I_i(\geq t)\big]$	The downstream can only use upstream information available at time t and earlier.
Jointly optimizable objectives	G	$\exists L_f = \alpha L_i + \beta L_j + \gamma L_{\mathrm{joint}}$	A joint objective can be defined and optimized.
Alignable granularity	D	$\mathrm{Freq}_i \sim \mathrm{Freq}_j \ \wedge\ A_i \cap A_j \neq \varnothing$	Data frequency and scope can be aligned.
Unifiable evaluation	E	$\mathcal{E}_{ij} = \mathcal{E}_i \cup \mathcal{E}_j \cup \mathcal{E}_{\mathrm{joint}}$	Evaluation can be unified on the same test set.
Causal consistency	C	No look-ahead bias $\wedge$ No information leakage $\wedge$ Satisfies domain constraints	Both training and inference satisfy causality.
Feasible resources	R	$\mathrm{Cost}(P_f) \leq \mathrm{Budget}$	Computational, time, storage, and other resources are within budget.

C Analysis of the Seven-Dimensional Determination Conditions for Six Fusion Directions

The scoring criteria are: 2 points for easy to satisfy, 1 point for conditionally satisfiable, and 0 points for difficult to satisfy. Detail descriptions are in Table 9.

Table 10: Item-by-item analysis of the seven dimensions for the six fusion directions.

Dim.	CFM→CTD	CFM→LTD	LFM→CTD	LFM→LTD	CTD→CFM	LTD→LFM
Interface Bridgeable	CFM outputs a factor pool; CTD receives factor value vectors; the interface is direct and the bridge mapping is simple. **[2pts]**	CFM outputs factor values; LTD must convert factors into LLM-understandable input; the bridge mapping requires design. **[2pts]**	LFM outputs factor expressions and values; CTD receives numerical vectors; direct interface. **[2pts]**	LFM outputs factor expressions and explanations; LTD receives factor information as LLM context; bridgeable. **[2pts]**	CTD outputs actions or weights; CFM needs factor expressions or values; not direct; complex bridging required. **[1pts]**	LTD outputs decisions and rationales; LFM needs factor-generation prompts; not direct; complex bridging required. **[1pts]**
Temporal No-leakage	Factor computation uses only historical data; decisions proceed in time order; no look-ahead bias. **[2pts]**	Must ensure LLM prompts contain no future information; factor computation and LLM decisions must be time-aligned. **[1pts]**	Factor computation uses historical data; decisions proceed in time; no look-ahead bias. **[2pts]**	The LLM context window must be strictly restricted to prevent future-information leakage. **[1pts]**	Strategy feedback must be back-tracked in time; aligning feedback with factor generation is difficult. **[1pts]**	Aligning LLM decision feedback with factor generation in time is difficult and prone to look-ahead bias. **[0pts]**
Objective Jointable	CFM optimizes IC and CTD optimizes Sharpe; can be coordinated via staged or joint optimization. **[2pts]**	CFM optimizes IC; LTD optimizes Sharpe and interpretability; the joint objective needs extra alignment terms. **[1pts]**	LFM optimizes IC and semantic consistency; CTD optimizes Sharpe; the joint objective needs coordination. **[1pts]**	Both involve LLMs; the joint objective must weigh factor quality and strategy performance. **[1pts]**	CTD optimizes Sharpe and CFM optimizes IC; objectives are inconsistent; the joint objective is hard to define. **[0pts]**	LTD optimizes Sharpe and interpretability; LFM optimizes IC and semantic consistency; large objective differences. **[0pts]**
Granularity Alignable	Both are daily frequency with a consistent asset universe; directly alignable. **[2pts]**	Frequency and asset universe alignable. **[2pts]**	Frequency and asset universe alignable. **[2pts]**	Frequency and asset universe alignable. **[2pts]**	Frequency and asset universe alignable. **[1pts]**	Frequency and asset universe alignable. **[1pts]**
Evaluation Unifiable	IC and strategy metrics can be evaluated uniformly under the same backtest framework. **[2pts]**	IC and LLM strategy metrics can be evaluated in one framework, but fusion metrics must be additionally defined. **[1pts]**	IC and strategy metrics can be evaluated uniformly. **[1pts]**	Evaluable in one framework, but fusion metrics must be defined. **[1pts]**	Strategy and factor metrics evaluable together, but feedback signals are noisy. **[1pts]**	Evaluable in one framework, but feedback signals are noisy. **[1pts]**
Causal Consistency	Factor computation and decisions use only historical information; causally consistent. **[2pts]**	Must guard against LLM hallucination and look-ahead bias; strict verification required. **[1pts]**	Must ensure LFM factor generation used no future information; causally consistent. **[1pts]**	Must guard against LLM hallucination and look-ahead bias; strict verification required. **[1pts]**	Strategy feedback may introduce future information; causal consistency hard to guarantee. **[0pts]**	Strategy feedback may introduce future information; causal consistency hard to guarantee. **[0pts]**
Resource Feasible	Conventional models have low compute cost and small resource requirements. **[2pts]**	LLM token costs are relatively high; moderate resource requirements. **[1pts]**	LFM requires LLM calls and CTD is cheap; total resources moderate. **[1pts]**	Both sides require LLM calls; highest token cost; large resource requirements. **[0pts]**	Conventional models are cheap; resources feasible. **[1pts]**	Both sides require LLM calls; highest token cost. **[0pts]**
Total	**14, High feasibility**	**9, Medium**	**10, Medium-High**	**8, Medium-Low**	**5, Low**	**3, Very Low**

D Factor Injection Methods and Their Characteristics

Table 11: Factor injection methods and their characteristics.

No.	Injection Method	Approach	Example	Pros	Cons	Applicable Scenarios
1	Structured Value Injection (VAL)	Structured prompt text: recent-values table under anonymized aliases (F001–F005) + verbatim formula with orientation sign + frozen directional LGBM forecast; leakage-safe.	`| 06-20 | F001 | +1.21 | -0.43 | F001: Mom_20d sign=+1 LGBM h=1: +1.21%`	Lossless, precise, reproducible (frozen payloads); aliases block name priming.	Numbers less salient to LLMs; token cost grows with factor count; without semantics, agents may override the forecast without evidence and over-trade.	Few factors ($\leq$5); LTD (prompt injection) or CTD (state augmentation); payload base for EXP/GOV
2	Structured Value with Semantic Explanation (EXP)	VAL payload + one textual explanation per factor, from the miner's native description or a frozen post-hoc LLM text.	`F001=+1.21 desc: ''20-day momentum; positive values indicate an upward trend''`	Precision + interpretability; high LLM utilization; explanation source controlled across arms.	Costlier than VAL; post-hoc text can be generic; small factor sets only; does not stop unjustified overrides.	LTD with few factors pursuing fusion gains
3	Value Injection with Decision Governor (GOV)	VAL uses two-stage constraints: announced before decisions, then enforce forecast-consistent direction, strength-scaled positions (full $\geq$ 100 bp), and Hold below 20 bp; overrides are half-sized, while Holds determined by portfolio manager remain unchanged.	`Buy -> Hold (idle band, |f|<20bp) OVERRIDE: ''CPI 8.2%...'' -> kept at 0.5x size`	Deterministic, auditable; curbs unjustified overrides and over-trading while keeping justified judgment; identical payload to VAL $\Rightarrow$ gains attributable to the constraint.	Needs a directional forecast panel; three hyper-parameters; throttle invisible in single-stock P&L; may cap idiosyncratic LLM views.	LTD with directional forecast; default for the main fusion results; constraint arm of the 2×2 ablation

E Calculation Formulas for the Metrics of the EDD Evaluation System

This appendix provides the specific calculation formulas for each metric in the three-layer evaluation system described in Section 3.1. All metrics are calculated under a unified backtesting framework to ensure cross-paradigm comparability.

Table 12: Calculation formulas of evaluation indicators (I): common performance layer.

Evaluation Level		Calculation Formula of Evaluation Indicators
Common Layer	Performance	**Annualized Return (ARR):** $ARR = \left[\prod_{t=1}^{T}(1+r_t)\right]^{252/T} - 1$, where r_t is the return on day t, and T is the total number of trading days in the backtest. **Sharpe Ratio (SR):** $SR = \frac{\bar{r} - r_f}{\sigma_r}\sqrt{252}$, where $\bar{r}$ is the average daily return, r_f is the risk-free rate (usually set to 0), and σ_r is the standard deviation of daily returns.

(continued on next page)

Table 12 – continued from previous page

Evaluation Level	Calculation Formula of Evaluation Indicators
	Maximum Drawdown (MDD): $MDD = \max_t \frac{\text{Peak}_t - \text{Trough}_t}{\text{Peak}_t}$, where Peak_t is the peak cumulative net value up to day t, and Trough_t is the subsequent trough.
	Calmar Ratio (Calmar): $Calmar = \frac{ARR}{MDD}$.
	Turnover: $Turnover = \frac{1}{T-1}\sum_{i=2}^{T}\frac{1}{2}\sum_{j=1}^{n}\left\|w_{i,j} - w_{i-1,j}\right\|$, where $w_{i,t}$ is the weight of asset i on day t, and N is the number of assets.

Table 13: Calculation formulas of evaluation indicators (II): intrinsic capability layer.

Evaluation Level	Calculation Formula of Evaluation Indicators
Intrinsic Capability Layer: Factor Mining Dimension (CFM, LFM)	**Information Coefficient (IC):** $\text{IC}_t = \text{Corr}(f_t, r_{t+1}), \quad \overline{IC} = \frac{1}{T}\sum_{t=1}^{T} IC_t,$ where f_t is the cross-sectional value of the factor on day t, and r_{t+1} is the next-day return.
	IC Information Ratio (ICIR): $ICIR = \frac{\overline{IC}}{\text{std}(IC)}$.
	Rank IC: $\text{RankIC}_t = \text{Spearman}(f_t, r_{t+1}), \quad \overline{\text{RankIC}} = \frac{1}{T}\sum_{t=1}^{T}\text{RankIC}_t.$
	Rank ICIR: $\text{RankICIR} = \overline{\text{RankIC}}/\sigma(\text{RankIC}_t)$.
	Factor diversity (average pairwise correlation): $\text{Div} = \frac{2}{K(K-1)}\sum_{i=1}^{K}\sum_{j=i+1}^{K}\left\|\rho(f_i, f_j)\right\|$, where K is the number of factors in the pool and ρ is the Pearson correlation between factor values. A smaller value indicates higher diversity.
	Factor novelty (maximum AST similarity to Alpha158): $\text{Nov} = 1 - \max_{\alpha \in \text{Alpha158}} s(f, \alpha)$, where $s(f, \alpha)$ is the maximum common subtree similarity (normalized to $[0, 1]$) between factor f and factor α in Alpha158. A larger value indicates higher novelty.
Intrinsic Capability Layer: Trading Decision Dimension (CTD, LTD)	**Sharpe Ratio** and **Maximum Drawdown**: same as the Common Performance Layer.
	Extreme-Regime Adaptability (excess return in bear/range-bound markets): $\text{Excess}_{\text{regime}} = \frac{1}{\lvert T_{\text{regime}}\rvert}\sum_{t \in T_{\text{regime}}}\left(r_t^{\text{strategy}} - r_t^{\text{benchmark}}\right)$, where T_{regime} is the set of bear- or range-bound-market days, r_t^{strategy} is the daily strategy return, and $r_t^{\text{benchmark}}$ is the daily benchmark return.
	Decision Frequency Stability: $\text{Stab}_{freq} = 1 - \frac{\sigma_{\text{freq}}(\text{trade_count})}{\max\ \sigma}$, where trade_count is the number of trades per period, σ is the standard deviation, and the denominator is a normalizing factor. The coefficient of variation of the trade count can also be used directly.

(continued on next page)

Table 13 – continued from previous page

Evaluation Level	Calculation Formula of Evaluation Indicators
Intrinsic Capability Layer: Traditional Model Technical Dimension (CFM, CTD)	**Convergence Speed:** $\mathrm{Conv}_{speed} = N_{target}$, where N_{target} is the number of iterations or training time required to reach the target performance (e.g., validation set IC or Sharpe reaching a preset threshold).
	Convergence Stability: $\mathrm{Conv}_{stab} = \sigma_{runs}\big(\{\mathrm{Conv}_{speed}\}\big)$, i.e., the standard deviation of convergence speed across multiple independent runs.
	Model Complexity: Factor expression: $Complexity = \lvert V(T_f)\rvert$, i.e., the number of AST nodes; Strategy network: $Complexity =$ number of parameters.
Intrinsic Capability Layer: Large Model Technical Dimension (LFM, LTD) **Code Generation and Executability Quality**	**Syntax pass rate:** $\dfrac{\#\text{ syntactically correct}}{\#\text{ total generated}}$
	Semantic executability rate: $\dfrac{\#\text{ executable}}{\#\text{ syntactically correct}}$
	Semantic validity rate: $\dfrac{\#\text{ produce valid signals}}{\#\text{ executable}}$
	First-pass rate: $\dfrac{\#\text{ pass on first generation}}{\#\text{ total tasks}}$
(LFM, LTD) **Inference Efficiency and Cost Control**	**Effective output per token:** $\dfrac{\Delta\mathrm{Sharpe}}{N_{token}}$
	Effective output per dollar: $\dfrac{\Delta\mathrm{Sharpe}}{\mathrm{CostUSD}}$
	End-to-end latency: $\dfrac{1}{T}\sum_{i=1}^{T}\Delta t_i$, i.e., the average time per decision.
(LFM, LTD) **Context Compression Ratio**	**Input compression ratio:** original input length / compressed input length.
	Output conciseness: valid information / total output length.
	Information density: valid information / total tokens.

Table 14: Calculation formulas of evaluation indicators (III): fusion evaluation layer

Evaluation Level	Calculation Formula of Evaluation Indicators
Fusion Evaluation Layer	**ARR gain:** $\Delta\mathrm{ARR} = \mathrm{ARR}_{\mathrm{FM+TD}} - \mathrm{ARR}_{\mathrm{TD}}$.
	Sharpe gain: $\Delta\mathrm{Sharpe} = \mathrm{Sharpe}_{\mathrm{FM+TD}} - \mathrm{Sharpe}_{\mathrm{TD}}$.
	Factor source marginal contribution (MC): $\mathrm{MC} = \dfrac{\Delta\mathrm{Sharpe}}{\Delta\mathrm{Sharpe}_{\mathrm{FM}}}$, where $\Delta\mathrm{Sharpe}_{\mathrm{FM}} = \mathrm{Sharpe}_{\mathrm{FM}} - \mathrm{Sharpe}_{\text{no-factor baseline}}$.
	Factor fidelity decay (FFD): $\mathrm{FFD} = 1 - \rho\big(\mathrm{sign}(f), \mathrm{action}\big)$, where ρ is the correlation coefficient between the factor signal direction and the trading action.

F SCORING AND SCREENING OF PARADIGM INSTANCE FUSION COMBINATIONS

F.1 SCORE RANKING OF PARADIGM INSTANCE COMBINATIONS

Following Step 2 of TriSift, fusion combination scoring. For each candidate fusion combination (F, T), where F is the factor mining instance and T is the trading decision instance, the screening score is defined as: $Score(F, T) = \alpha R_F + \beta R_T$ where R_F is the internal composite ranking score of the factor mining instance, R_T is the internal composite ranking score of the trading decision instance, and the weight parameters are $\alpha = 0.5$, $\beta = 0.5$.

Based on the above formula, the scores of each paradigm instance combination are calculated as shown in Table 15.

Table 15: Score Rankings of All Paradigm Instance Combinations

Rank	Fusion Combination	R_F	R_T	Score
1	AlphaGen + TradingAgents	1.25	1.56	1.405
2	AlphaGen + FinRL-PPO	1.25	2.00	1.625
3	AlphaQCM + TradingAgents	1.75	1.56	1.655
4	AlphaQCM + FinRL-PPO	1.75	2.00	1.875
5	AlphaProbe + TradingAgents	2.429	1.56	1.995
6	AlphaGen + FinRL-A2C	1.25	2.78	2.015
7	AlphaAgent + TradingAgents	2.786	1.56	2.173
8	AlphaProbe + FinRL-PPO	2.429	2.00	2.215
9	AlphaQCM + FinRL-A2C	1.75	2.78	2.265
10	QuantaAlpha + TradingAgents	3.000	1.56	2.280
11	RD-Agent + TradingAgents	3.214	1.56	2.387
12	AlphaAgent + FinRL-PPO	2.786	2.00	2.393
13	AlphaGen + FinMem	1.25	3.67	2.460
14	QuantaAlpha + FinRL-PPO	3.000	2.00	2.500
15	FactorMiner + TradingAgents	3.571	1.56	2.566
16	AlphaProbe + FinRL-A2C	2.429	2.78	2.605
17	RD-Agent + FinRL-PPO	3.214	2.00	2.607
18	AlphaQCM + FinMem	1.75	3.67	2.710
19	AlphaAgent + FinRL-A2C	2.786	2.78	2.783
20	FactorMiner + FinRL-PPO	3.571	2.00	2.786
21	QuantaAlpha + FinRL-A2C	3.000	2.78	2.890
22	RD-Agent + FinRL-A2C	3.214	2.78	2.997
23	AlphaProbe + FinMem	2.429	3.67	3.050
24	FactorMiner + FinRL-A2C	3.571	2.78	3.176
25	AlphaAgent + FinMem	2.786	3.67	3.228
26	QuantaAlpha + FinMem	3.000	3.67	3.335
27	RD-Agent + FinMem	3.214	3.67	3.442
28	FactorMiner + FinMem	3.571	3.67	3.621

F.2 SCREENING OF CANDIDATE PARADIGM INSTANCES FOR FUSION

Following the coverage-balance adjustment requirement in Step 3 of TriSift, for the above candidate combinations, two top-ranked paradigm instances from each paradigm group can be considered as candidates. According to the results in Table 15, the top two for CFM are AlphaGen and AlphaQCM; for LFM, AlphaProbe and AlphaAgent; for CTD, FinRL-PPO and FinRL-A2C; and for LTD, TradingAgents and FinMem. This yields 16 paradigm combinations $(2 + 2) \times (2 + 2)$. However, since FinRL-PPO and FinRL-A2C belong to the same series, FinRL-PPO is selected as the representative to participate in the combinations. The finally screened candidate paradigm instances for fusion are: upstream factor mining paradigm instances: AlphaGen, AlphaQCM, AlphaProbe,

AlphaAgent; downstream trading decision paradigm instances: TradingAgents, FinMem, FinRL-PPO. These can form $4 \times 3 = 12$ paradigm fusion combinations.

G Experimental Setup Description

All experiments were completed on the unified evaluation platform ParaFusionBench: each method is connected to the same driver through a standard adapter, and the platform uniformly handles data loading, sandbox isolation, backtest execution, and metric computation. This appendix provides the complete setup of the unified protocol for five U.S. stocks (US5); the numerical values in the tables in the main text are not repeated here.

G.1 Evaluation Subjects

The four paradigms are cross-divided along "task mainline $\times$ technical route": CFM, LFM, CTD, LTD. The methods under evaluation are selected according to three criteria: publicly published, open-source and reproducible, and community representative (a few influential methods not formally published, such as TradingAgents, are also included).

Table 16: Evaluation subjects across the four paradigms

Paradigm	**Method**
Market Baseline	Buy & Hold (Equal-weighted holding of the five stocks)
Traditional Factor Library	alpha20, alpha158
CFM	AlphaGen, AlphaQCM
LFM	RD-Agent, QuantaAlpha, AlphaAgent, AlphaProbe, FactorMiner
CTD	FinRL (Two algorithms, A2C and PPO, with PPO as the representative in the main text)
LTD	TradingAgents, FinMem

G.2 Data, Instruments, and Time Windows

Test instruments: AAPL, GOOGL, MSFT, NVDA, TSLA (following the common choices of existing benchmarks such as TradingAgents and AlphaForgeBench; all are highly liquid instruments with a wide range of volatility characteristics).

Data sources: daily market data from Yahoo Finance; news from Alpaca/Finnhub historical archives (pre-downloaded); financial reports from SEC EDGAR; the S&P 500 index and historical constituent lists for market regime classification and factor mining universe construction (point-in-time filtering to avoid survivorship bias).

Time windows: training 2019-01-01 to 2023-12-31, validation 2024-01-01 to 2024-12-31, and testing the full year 2025 (daily frequency, approximately 250 decision days per stock). The validation window is used for model selection and factor screening; the test window remains blind until the final unified backtest.

Protocol exceptions: (1) As per its method design, FinMem has a memory formation period from 2023-01-01 to 2024-12-31; (2) Fusion experiments use the five stocks from 2025-01-01 to 06-30 (first half; 128 decision days per stock for TradingAgents and 121 for FinMem), and ablation experiments are further narrowed to the same-period window for GOOGL only (121 decision days).

Market regime classification: for the S&P 500 index, a 60-day cumulative return $> +5\%$ denotes a bull market, $< -5\%$ a bear market, and in between a range-bound market; days with 20-day rolling volatility above the median are high-volatility days. The sample sizes for the four categories are 121/37/92/125, respectively.

G.3 Backtesting Protocol

See Table 17 for detailed backtesting protocol settings.

Table 17: Backtesting protocol settings

Item	Setting
Frequency and price	Daily frequency; settled at same-day closing price
Portfolio semantics	Score-to-weight dollar-neutral long–short portfolio: scores > 0 are long and < 0 are short, weighted proportionally to scores within each side (long and short each ± 1), rebalanced daily; when all scores share the same sign, it degenerates to a one-sided portfolio
Transaction costs	U.S. equities: 1 bp one-way $\times$ turnover; no price limit restrictions (the U.S. market has no fixed price limits)
LLM decision execution	A unified executor maps decisions to weights, then deducts costs and settles (TradingAgents $w = \text{score}/2$; FinMem $w = \text{action} \in \{-1, 0, +1\}$)
Baseline	Buy & Hold, with the same cost basis

G.4 Model and Budget Configuration

- **Unified configuration for large models:** deepseek-flash (thinking disabled), temperature $= 0$, provider random seed bound to the run seed. This is a unified, cost-controlled setting for all methods and does not represent the performance upper bound of any method.
- **Factor mining budget (CFM/LFM):** 3 hours of wall-clock time + unified top-20 factor screening; at the deadline, the accumulated factors enter blind backtesting.
- **Trading decision (LTD):** no mining budget; daily decisions throughout the test window; FinRL (CTD): training budget of 100k steps, checking validation Sharpe ≥ 1.0 every 10k steps.
- **Return prediction synthesis:** factor values from each factor-track method are uniformly synthesized into return prediction scores via the same LightGBM configuration (hyperparameters tuned on 5-day forward return labels; no separate tuning for any method). The prediction horizon is uniformly $h = 5$, which only determines the alignment window for training labels and information coefficients and does not change the daily rebalancing frequency; the injection prediction panel for fusion experiments separately uses $h = 1$. The information coefficients for $h = 1$ and $h = 5$ refer to different forward returns, so their values are not directly comparable.
- **Traditional factor library protocol:** Alpha20/Alpha158 follow the same protocol as CFM/LFM—full cross-section (S&P 500) training and scoring, reporting only the five-stock portfolio; Rank IC is also calculated on the full cross-section (the cross-sectional rank correlation of five stocks has no statistical significance).
- **FinRL special case:** the policy dimension of end-to-end reinforcement learning is tied to trading assets; therefore, training/validation/testing are completed within the five-stock environment, and full cross-section data are used only for market regime classification.

G.5 Paradigm Fusion Setup

- **Direction and coverage:** factor mining $\rightarrow$ trading decision, covering all four combination types of {CFM, LFM} $\times$ {FinRL-PPO, TradingAgents, FinMem}; fusion instances are determined by hard-threshold filtering and full-dimension rank-weighted screening, and the fusion budget is 4 factor sources $\times$ 3 consumer ends, totaling 12 combinations.
- **Injection channels:** for large-model consumer ends, prompt injection; for FinRL, state augmentation (directional factor values are incorporated into the environment's technical indicator channel).
- **Three levels of injection methods**:
  - VAL = factor value table + verbatim formula + directional prediction;
  - EXP = VAL+ economic logic explanation of factors;
  - GOV = VAL+ decision constraint layer (direction reversal must be justified; justified reversal halves the position; position scales with prediction strength; weak-prediction dead zone defaults to holding).

  The main table uniformly uses GOV.

- **Factor selection and freezing:** successfully executed factors from each source are used to calculate full-cross-section 1-day Rank IC in the 2024 validation window $\rightarrow$ remove bad factors $\rightarrow$ greedily deduplicate in descending order of $|IC|$ (remove pairs with Spearman $|\rho| > 0.7$) $\rightarrow$ orient by the sign of the information coefficient $\rightarrow$ freeze top-5 as the list; no re-selection is performed during fusion runs. Factor values are standardized daily across the full cross-section and multiplied by the direction sign; the prediction panel is injected together with the factor list, with strict alignment (missing values trigger an error).
- **Runtime preconditions:** the same agent configuration, the same large model, the same decision calendar, the same frozen factor list, and the same prediction panel.

G.6 Ablation Experiment Setup

- **Carrier:** AlphaProbe $\times$ FinMem, GOOGL single stock from 2025-01-01 to 06-30, seed $= 123$; all ablation arms share the same model configuration, decision calendar, and 2023–2024 hierarchical memory (not injected during training; memories are identical across arms). The baseline is the no-injection FinMem on the same calendar slice; the backtesting protocol is consistent with the FinMem row.
- **Injection method (2×2):** two payload levels (VAL / EXP) $\times$ two constraint levels (none / GOV), giving four arms: VAL, EXP, GOV, and EXP + GOV.
- **Number of factors (GOV fixed):** inject 1/5/10 cross-type independent factors—the factor pool is clustered by pairwise correlation, and the best performer in the validation window is taken from each of the top K clusters with the lowest correlation; at the beginning of each month, factor direction validity is reviewed using a rolling window, and those that continuously fail are automatically deactivated; the three arms share the same prediction panel, with the only difference being the number of injected factors.
- **Factor type (GOV fixed):** inject the top-5 within each of the three categories—momentum, volatility, and price–volume (deterministic grouping by formula form, selected and frozen in the 2024 validation window); the mixed type is the GOV row.

FinRL is end-to-end reinforcement learning (factors can only be incorporated into the state vector) and is not applicable to injection method ablation.

H Supplementary Experimental Results

This appendix collects supplementary experimental figures of ParaFusionBench, including the fusion gain across the twelve candidate combinations, the consumer-side decomposition of fusion gains, the per-ticker cumulative return paths of the fusion arms on AAPL, GOOGL, MSFT, NVDA, and TSLA, the decoupling between cross-sectional factor quality and realized P&L, the ablations on the number and type of injected factors, and the overall performance of the evaluated paradigm instances (RQ1).

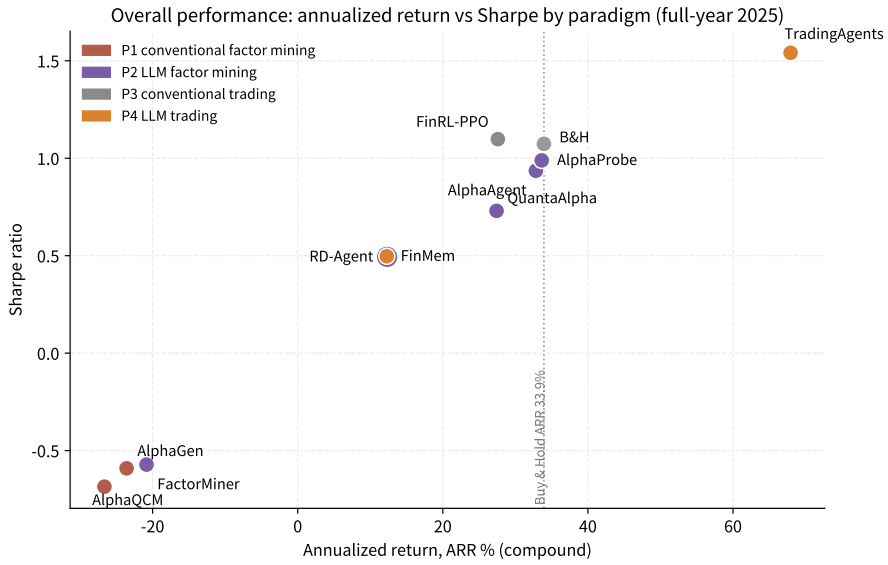


Figure 3: Overall performance of the evaluated paradigm instances over the full year 2025 (RQ1): annualized return (compound) versus Sharpe ratio, colored by paradigm (CFM, LFM, CTD, LTD); the dotted vertical line marks the Buy & Hold annualized return (33.9%).

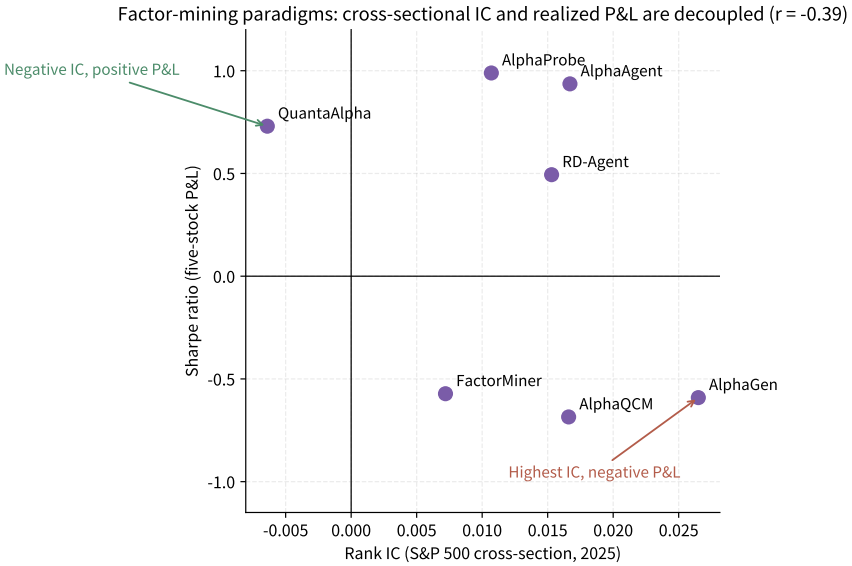


Figure 4: Decoupling between cross-sectional Rank IC and realized five-stock P&L across factor-mining paradigms.

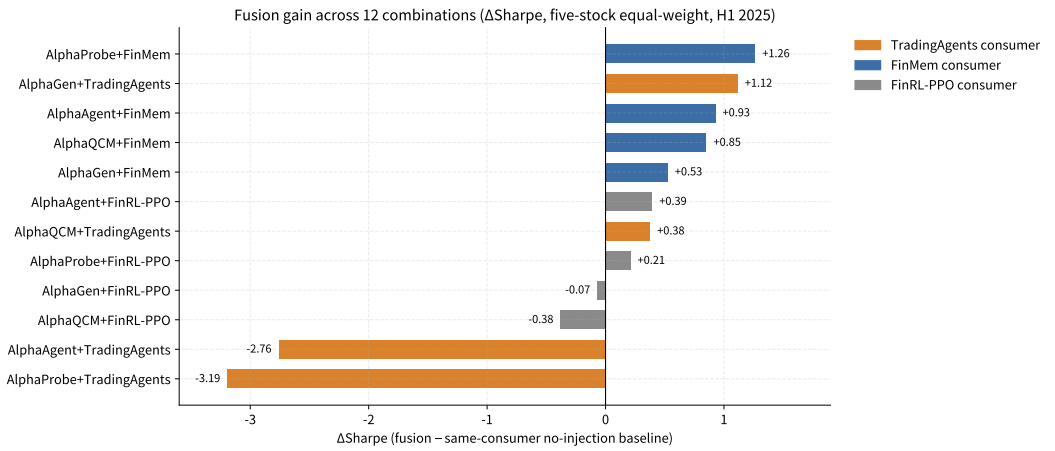


Figure 5: Fusion gain in Sharpe across the twelve candidate combinations (five-stock equal-weight, 2025 H1). ΔSharpe is measured against the same-consumer no-injection baseline.

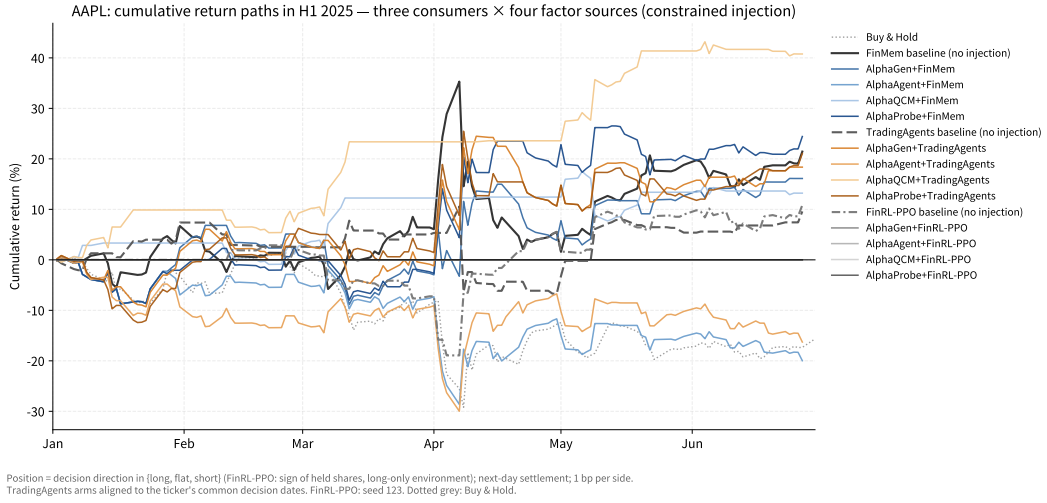


Figure 6: Cumulative return paths on AAPL over 2025 H1 under the GOV injection protocol: the twelve fusion arms (four factor sources × three consumers), the three no-injection consumer baselines, and Buy & Hold (dotted grey).

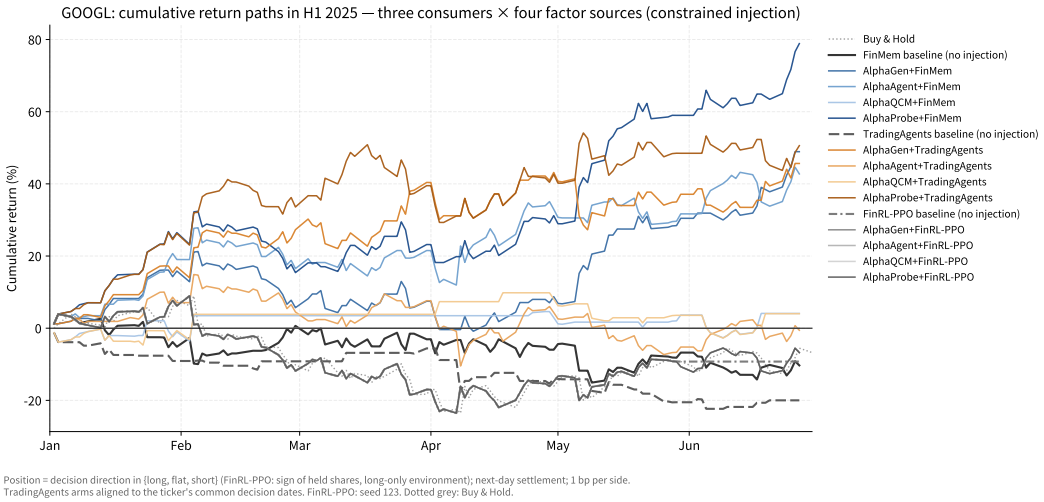


Figure 7: Cumulative return paths on GOOGL over 2025 H1 under the GOV injection protocol: the twelve fusion arms (four factor sources × three consumers), the three no-injection consumer baselines, and Buy & Hold (dotted grey).

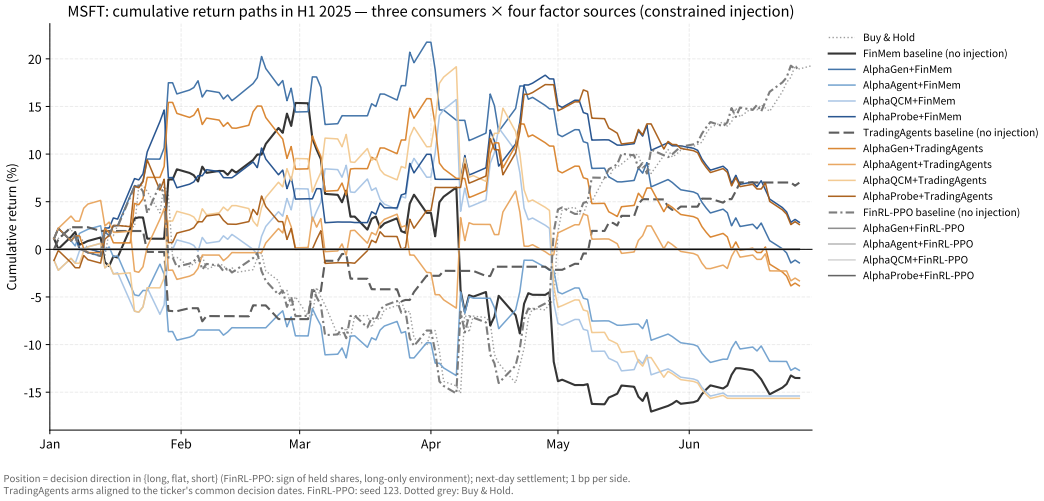


Figure 8: Cumulative return paths on MSFT over 2025 H1 under the GOV injection protocol: the twelve fusion arms (four factor sources × three consumers), the three no-injection consumer baselines, and Buy & Hold (dotted grey).

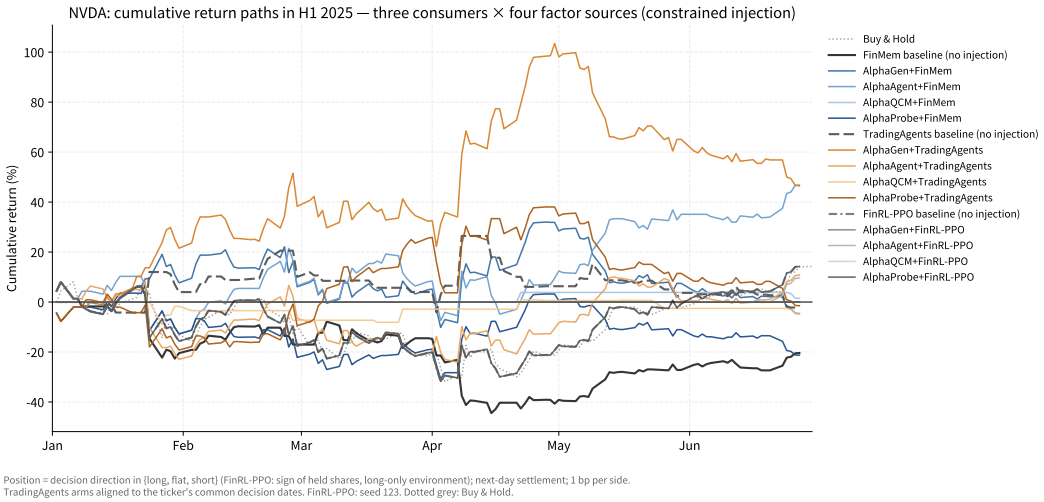


Figure 9: Cumulative return paths on NVDA over 2025 H1 under the GOV injection protocol: the twelve fusion arms (four factor sources × three consumers), the three no-injection consumer baselines, and Buy & Hold (dotted grey).

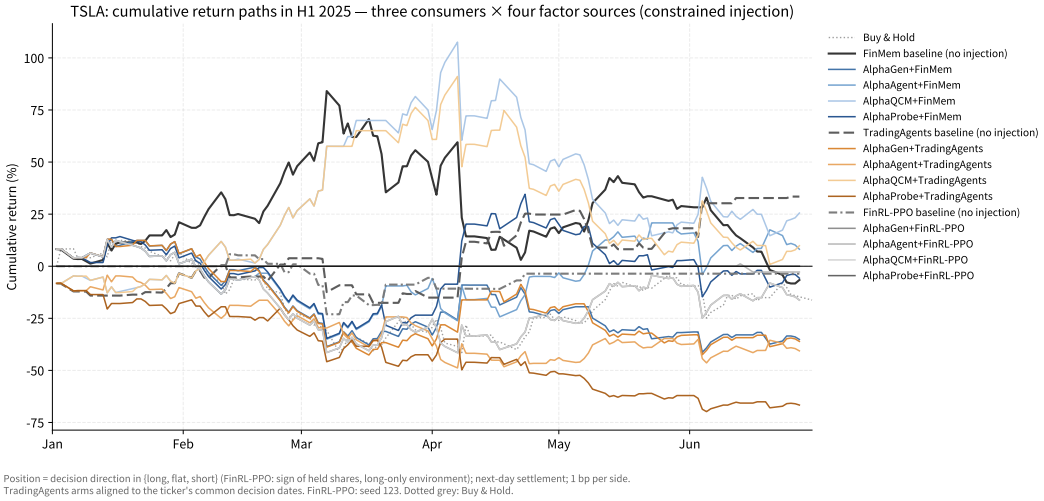


Figure 10: Cumulative return paths on TSLA over 2025 H1 under the GOV injection protocol: the twelve fusion arms (four factor sources × three consumers), the three no-injection consumer baselines, and Buy & Hold (dotted grey).

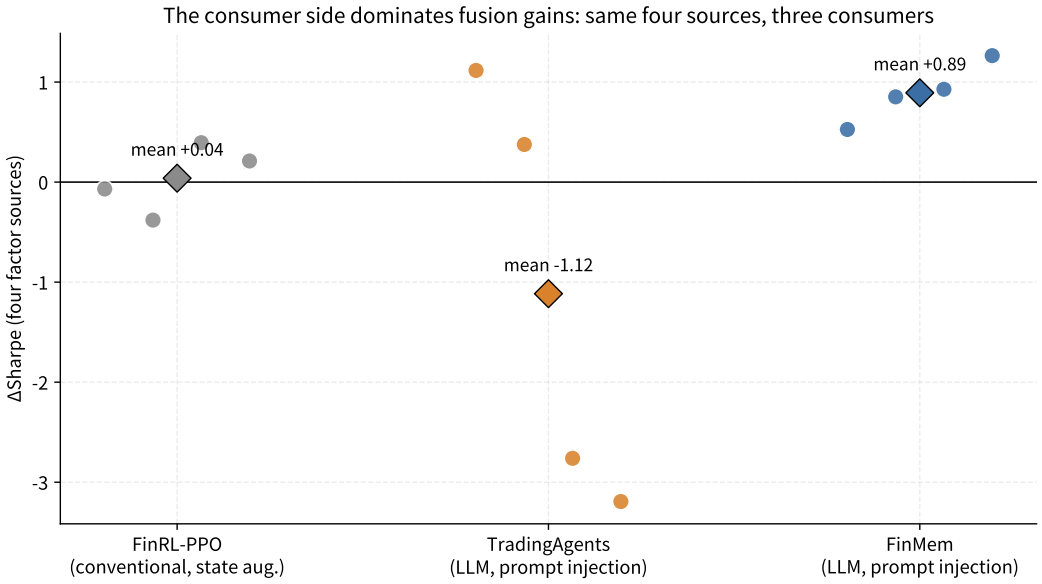


Figure 11: Consumer-side dominance of fusion gains: ΔSharpe of the same four factor sources across the three consumers; diamonds mark consumer means.

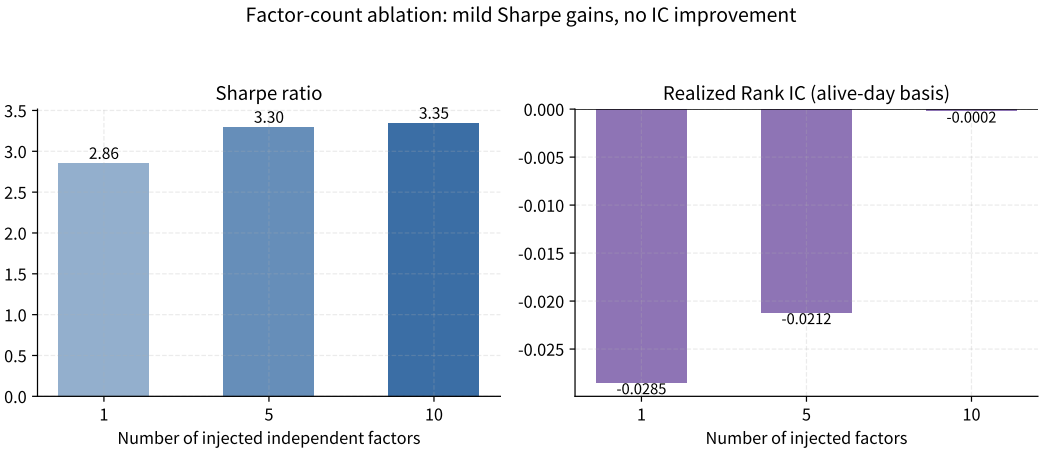


Figure 12: Effect of the number of injected independent factors on the Sharpe ratio and the realized Rank IC.

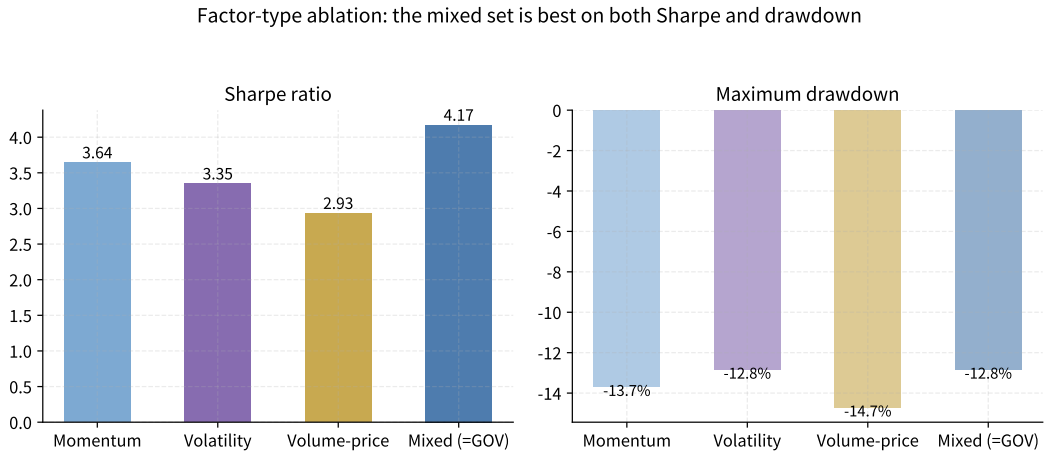


Figure 13: Effect of the injected factor type on the Sharpe ratio and the maximum drawdown of the fused strategy.