PPD-SOTA: RECASTING STATE-OF-THE-ART EVALUATION FOR AI-DRIVEN FINANCIAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-driven financial forecasting has advanced rapidly, but state-of-the-art evaluation still relies heavily on metric wins over selected baselines. In financial forecasting, however, predictive performance must pass through portfolio construction and deployment conditions before it translates into practical value, making a single leaderboard rank difficult to interpret. We introduce PPD-SOTA, an evaluation framework that separates forward model selection from the scope of a SOTA claim. For deployment, PPD-SOTA evaluates forecasting models through **Prediction, Portfolio, and Deployability (PPD)**. For comparison, it tests a comparator set using practical margins and paired simultaneous inference. On a unified benchmark of recent stock-forecasting methods, PPD achieves lower mean regret than conventional metrics-win selection and remains effective across changing conditions, while superiority support narrows as the comparator set expands and no model clears the full set. These results motivate evaluating financial forecasting models by both forward deployment performance and the scope of superiority supported by the evidence, rather than by isolated metric wins alone.

1 INTRODUCTION

Financial forecasting studies often summarize model performance with a leaderboard built from prediction and portfolio metrics. Such rankings are convenient, but they compress several distinct evaluation choices into a single ordering. Recent financial AI studies commonly compare models using these metrics, with baseline wins often serving as the basis for SOTA claims (Zhao et al., 2023; Li et al., 2024b; Fan & Shen, 2024; Duan et al., 2025). The resulting ranking, however, can change with the prediction target, evaluation window, backtest construction and so on. Repeated backtest selection can also introduce optimism (White, 2000; Bailey & López de Prado, 2014), while stochastic training and dependent returns complicate uncertainty estimation (Henderson et al., 2018; Agarwal et al., 2021; Politis & Romano, 1992). The largest point estimate alone can therefore determine neither which model should be deployed next nor whether the evidence supports a broader superiority claim.

These decisions serve different purposes. A quantitative researcher needs a defensible next-period choice together with its strengths and weaknesses. Authors and reviewers in financial AI face a different question: whether an observed advantage supports a local result, a broader SOTA claim, or no resolved claim. PPD-SOTA therefore separates the model chosen for deployment from the scope of the superiority claim supported by the evidence.

Ranking instability makes same-window fit insufficient for model selection; the chosen model must be assessed in multiple out-of-sample test periods. Forecasts also have no real investment utility until they are mapped to trading strategies with explicit rules for costs, cash, and execution prices. Beating one baseline on a point estimate does not establish superiority over a comparator family. PPD-SOTA therefore combines forward PPD selection, execution-consistent account replay, and simultaneous comparison against the declared competitors. PPD-SOTA separates forward deployment selection from scientific superiority qualification, treating them as distinct decisions with different evidence requirements.

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.

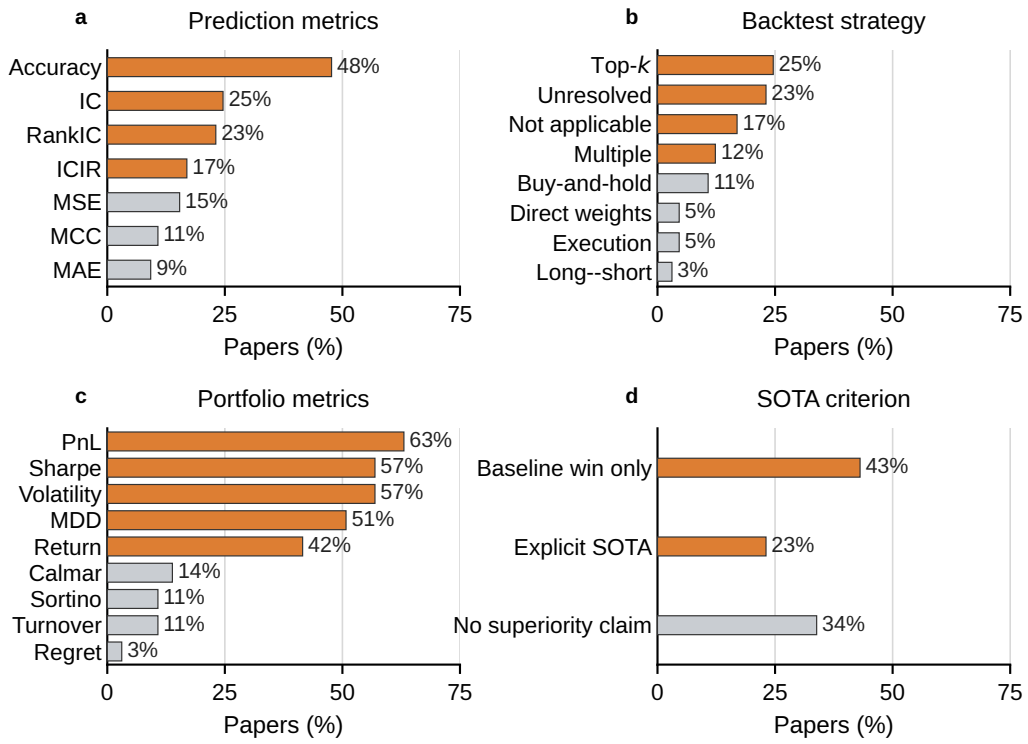

Figure 1: Heterogeneity of evaluation practices in recent financial AI literature ($n = 65$). Panels a and c count every recovered metric, so categories can overlap; panel b assigns each paper one strategy state, with unresolved evidence distinct from a strategy that is not applicable. Panel d is an exhaustive, mutually exclusive partition: baseline win only, explicit SOTA, or no superiority claim.

Figure 1 documents substantial variation in metrics, signal-to-portfolio mappings, reported outcomes, and superiority criteria across 65 recent papers. We evaluate seven recent forecasting models under a common protocol that separates forward deployment selection from superiority claims. PPD achieves lower average next-quarter regret than literature-style metrics-win selection, while calibrated comparison against the full six-opponent set yields no clear winner. The two results point to the same conclusion: selecting a model for deployment and establishing the scope of its superiority require different evidence.

2 Related Work

Financial forecasting models. Financial forecasting research builds predictive signals in two broad ways. One line of work focuses on factor construction. Factor-mining methods search for explicit alpha expressions using neural search, reinforcement learning, or LLM agents (Shi et al., 2025; Yu et al., 2023; Tang et al., 2025), while factor models learn latent return signals without requiring explicit formulas (Duan et al., 2022; 2025; Yang et al., 2025). A second line of work changes the forecasting model itself, ranging from market-guided and cross-stock architectures (Li et al., 2024b; Fan & Shen, 2024; Xia et al., 2024; Cao et al., 2024; Li et al., 2025d) to adaptation and state-space models (Zhao et al., 2023; Zhang et al., 2024c; Mei et al., 2025; Shi et al., 2026). Because reported results come from different markets, targets, and evaluation protocols, we re-evaluate these methods under a common estimand rather than compare headline numbers across papers. Moreover, not all recent models provide public implementations, which limits controlled re-evaluation. We therefore focus on seven recent methods with author-released code and published evidence of competitive performance against multiple baselines, and re-evaluate them under a common estimand.

Financial benchmarks and evaluation. Qlib, FinRL-Meta, TradeMaster, QuantBench, and FinWorld improve reproducibility by standardizing different parts of financial AI research, from data and market environments to backtesting and evaluation (Yang et al., 2020; Liu et al., 2022; Sun et al.,

2023a; Wang et al., 2025; Zhang et al., 2026b). They compare systems mainly through performance measures, with some also incorporating reliability diagnostics. AlphaEval extends multidimensional evaluation to formula-alpha mining at the generated-alpha level (Ding et al., 2026). These efforts improve evaluation breadth but do not explicitly provide a complete-comparator, uncertainty-aware rule for determining the scope of a SOTA claim about forecasting models.

Statistical evaluation and reproducibility. Addressing this gap requires calibrated inference over a declared comparator set, not simply more evaluation metrics. Model confidence sets, equivalence tests, multiple-comparison correction, and block resampling provide complementary tools for this purpose (Hansen et al., 2011; Lakens, 2017; Holm, 1979; Politis & Romano, 1992). Model cards and datasheets further provide foundations for explicit and reconstructable declarations of evaluation scope (Mitchell et al., 2019; Gebru et al., 2021). Together, these lines of work provide useful statistical and reporting foundations, but do not by themselves define how SOTA claims should be scoped across a complete set of financial forecasting models. Our contribution therefore lies not in introducing another leaderboard metric, but in defining an evaluation protocol.

3 PPD-SOTA: Forward Selection and Qualified Superiority

3.1 Two decisions and three challenges

Metric-based ranking is natural in standard machine learning benchmarks, where predictions are evaluated directly against the target task, as in image classification. Financial forecasting, however, introduces an additional decision layer: predictions must be translated into positions and evaluated through a backtest with its own portfolio, execution, and cost assumptions. The recovered literature nevertheless reflects two common practices: some studies rely on metric-based rankings across prediction and portfolio outcomes, while others support SOTA claims through wins over selected baselines. These practices can obscure the source of apparent quality: high average cross-sectional correlation does not ensure that the investable tail earns higher subsequent returns.

Figure 2 makes this distinction concrete by tracing two evaluation paths from the same candidate model and baseline set. Conventional evaluation counts first-place results across metrics and uses the tally to support a SOTA claim or no claim. The PPD-SOTA operational path evaluates Prediction, Portfolio, and Deployability using evidence matched to each layer. The selected model is then compared with the full declared comparator set using practical margins and simultaneous inference to determine the scope of the superiority claim.

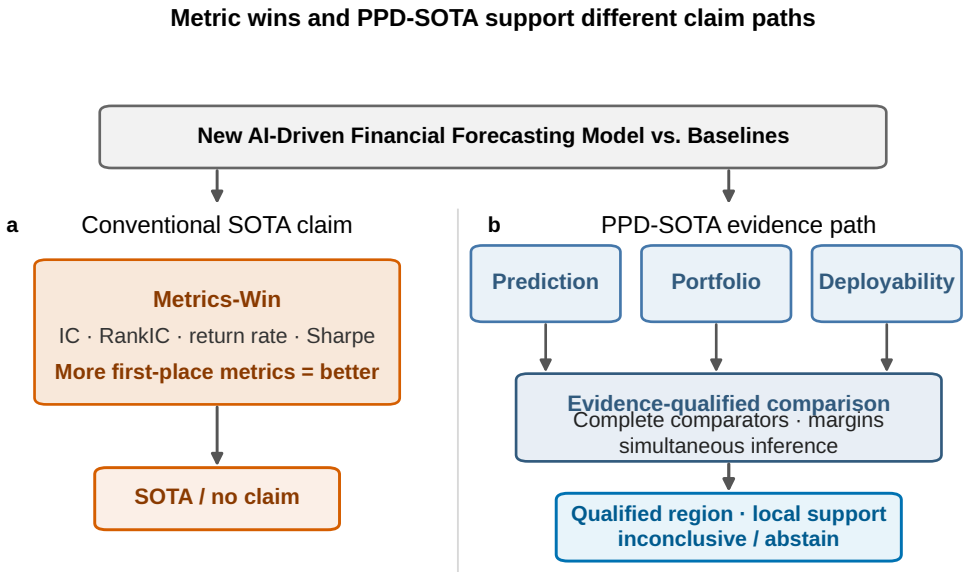

Figure 2: Two claim paths from the same model–baseline comparison.

3.2 FORWARD PPD MODEL SELECTION

For each quarterly decision t, PPD selects a model using only the trailing 12-month Test 1 window; the selected model is evaluated in the subsequent non-overlapping Test 2 quarter. Within Test 1, each metric is oriented so that larger values are better and converted to an event-internal percentile across the same seven models. P averages IC, RankIC, Top-10% spread, and IC positive ratio. Q averages annualized return, net Sharpe, maximum drawdown, gain concentration, and worst-environment Sharpe. The latter is the minimum net Sharpe over the complete $3 \times 2 \times 4$ Test-1 grid of holding fractions, rebalance schedules, and costs. D averages inverse turnover and operational reliability:

$$\mathrm{PPD} = f(P, Q, D). \tag{1}$$

The selected model is $\hat{f}_t = \arg\max_f \mathrm{PPD}_t(f)$ and is measured only in the next non-overlapping Test 2 quarter. Regret is the best available Test-2 net Sharpe minus that of the selected model. Test 2 never enters normalization, weights, or selection, so the target is the forward consequence of the Test-1 decision.

3.3 EXECUTION-CONSISTENT ACCOUNT REPLAY

At each rebalance date, model signals observed at close t are ranked cross-sectionally over the eligible universe and translated into long-only, equal-weight target positions for execution at next scheduled execution open $t+1$. The corresponding one-period realized return is $\mathrm{open}[t+2]/\mathrm{open}[t+1]-1$. If a new target lacks an executable open price, the intended allocation remains in cash. Existing positions are rebalanced only by the difference between current and target value, so turnover is measured from actual traded notional. We first compute gross returns and turnover, then apply transaction costs to the same trades. Algorithm 1 and Table 3 in the Appendix provide the trading protocol and parameter settings.

3.4 EVIDENCE-QUALIFIED SUPERIORITY REPORTING

A higher point estimate alone is not enough to support a superiority claim. Let e be a declared environment and $\Delta_{f,g}(e) = S_n(f, e) - S_n(g, e)$. A paired simultaneous interval $[L, U]$ and practical margin δ define

$$\mathsf{V}(e) = \begin{cases} \textsc{Superior}, & L > \delta, \\ \textsc{Inferior}, & U < -\delta, \\ \textsc{Equivalent}, & L > -\delta \ \wedge \ U < \delta, \\ \textsc{Inconclusive}, & \text{otherwise.} \end{cases} \tag{2}$$

Failure to establish superiority is not equivalence. We next ask where model f is supported as superior to every model in the declared comparator set $\mathcal{C}$. The qualified region is

$$\widehat{\mathcal{R}}(f > \mathcal{C}) = \{e : \min_{g \in \mathcal{C}} L_{f,g,e} > \delta\}. \tag{3}$$

A full-scope superiority claim requires this region to cover the declared scope.

Proposition 1 (Control of false qualified regions)**.** *If simultaneous intervals cover every frozen comparator–environment effect with probability at least* $1 - \alpha$*, then* $\Pr[\widehat{\mathcal{R}} \subseteq \mathcal{R}_{\mathrm{true}}] \geq 1 - \alpha$.

The intuition is simple: if an environment is falsely included, at least one comparator must have a true effect no larger than δ, which contradicts the lower-bound condition in Eq. 3 whenever joint coverage holds. The guarantee is conditional on valid intervals and leakage-free evaluation.

4 EXPERIMENTAL DESIGN

Models and data. The same-task candidate pool contains seven methods published in 2024–2025. We preserve the mechanism that distinguishes each method while holding the local features and next-open target fixed. The rolling split, seed fusion, portfolio construction, and costs are also shared. The resulting models are controlled adaptations rather than exact reproductions of headline experiments; their metadata and preserved mechanisms appear in Appendix Table 2.

Frozen evidence and budget. There are 28 analyzed prediction paths: seven models, two pools, and two seeds. Every portfolio and statistic reuses these predictions. The two-pool deployment grid covers four periods (2023, 2024, 2025, 2026H1), three holding fractions (5/10/20%), daily/weekly rebalancing, and 0/10/25/50-bp round-trip costs. We use no zero or forward fill for held-position returns. The grid tests whether a Test-1 model choice remains effective as portfolio size, rebalancing, and trading costs change. The ten selection dates span 2024Q1–2026Q2 in each available pool, giving 20 pool-events.

Baselines and controls. All selectors use the same candidates, dates, and Test-1 information. They also share the account specification and tie order. Metrics become direction-aligned percentiles across the seven models. Conventional Metrics-Win counts global wins on IC, RankIC, annualized return, net Sharpe, and inverse maximum drawdown, with four wins implementing the 80% SOTA threshold. We also include several additional selector baselines, with metric choices guided by the literature statistics summarized in Figure 1. Mean-rank-2Q gives equal weight to conventional Prediction and Portfolio. Tail-aware-2Q adds tail metrics, and Robust-Portfolio-2Q additionally includes the Test-1 worst environment. PPD uses Eq. 1. Max-IC and Max-Net-Sharpe are single-metric controls. Analytic uniform Random is the no-information reference, and future-aware Oracle is an unattainable upper bound. Quantities called IR in historical code implement $\sqrt{252}\,\bar{r}/s(r)$ and are labeled Sharpe ratios here.

Research questions and contracts. We ask four questions:

- **RQ1:** Why are conventional leaderboard winners unreliable?
- **RQ2:** Does PPD improve forward model selection?
- **RQ3:** Does one frozen PPD selection remain useful across 24 deployment environments?
- **RQ4:** Which superiority claims survive calibrated simultaneous inference against the complete declared comparator set?

Each contract declares its unit, inputs, statistic, and controls, including the scope of interpretation. RQ1–RQ3 evaluate operational selection; RQ4 evaluates SOTA qualification.

5 Results

5.1 RQ1: Do conventional metrics identify reliable future winners?

Table 1: Central seven-model benchmark. Quality columns are mean event-internal percentiles on a 0–100 scale; Portfolio includes Test-1 worst-environment Sharpe. Ranks average 20 dependent pool-events. Higher quality and lower rank are better. Reliability is constant, so Deployability is distinguished by turnover.

Model	Prediction	Portfolio	Deploy.	Mean-rank 2Q	PPD rank
MASTER	55.7	37.3	40.8	4.05	4.40
StockMixer	25.6	62.2	87.5	4.65	3.40
CI-STHPAN	64.4	63.3	25.8	2.50	3.40
MATCC	45.4	50.7	62.5	4.35	3.90
UMI	62.5	41.2	29.2	3.95	5.00
FinD3	49.2	83.2	95.8	2.75	1.60
StockCL-Trans.	47.2	12.2	8.3	5.75	6.30

Table 1 provides a profile-level view of the seven models before forward selection. Prediction, Portfolio, and Deployability strengths do not align uniformly across models, so a single aggregate rank can hide important trade-offs. CI-STHPAN is strongest in Prediction, FinD3 in Portfolio and Deployability, while StockMixer remains relatively deployable despite weaker Prediction. The table summarizes profiles over the 20 dependent decisions. Global superiority is evaluated separately.

Having shown that model strengths differ across Prediction, Portfolio, and Deployability, Figure 3 examines whether conventional single-metric shortcuts provide reliable forward signals. Panel a keeps seeds separate across 28 model–pool–seed paths. Mean IC and Top-10% spread are strongly associated overall ($\rho_s = .86$), but three positive-IC paths have nonpositive selected-tail spreads. Panel

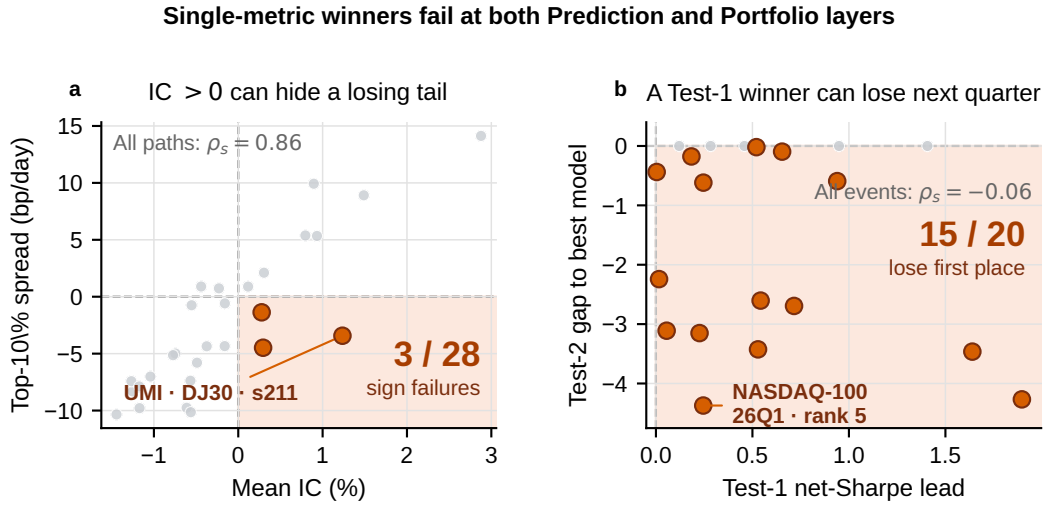

Figure 3: Two failure modes behind RQ1. (a) Positive average IC does not guarantee a positive investable-tail spread. (b) A Test-1 net-Sharpe winner need not retain next-quarter leadership under the fixed Top-10%, daily, 25-bp protocol. Panel a uses seed-level paths; panel b uses pool-quarter events.

b selects the Test-1 net-Sharpe winner in each of 20 pool-events; 15 lose first place during the next quarter. Its vertical coordinate is selected-minus-best Test-2 net Sharpe, so values below zero are forward reversals. Orange marks only these failure states; gray points provide context.

Answer to RQ1. The conventional single-metric winners are unreliable forward selectors in this benchmark. Positive IC does not always yield a profitable investable tail, and 15 of 20 Test-1 net-Sharpe winners lose first place next quarter. We therefore judge selection by forward transfer rather than same-window rank.

5.2 RQ2: Does PPD improve forward model selection?

Figure 4 reports the primary comparison. At each quarterly decision, PPD, Metrics-Win, and the displayed Max Sharpe and Max IC controls use only the trailing twelve-month Test 1 window. The selected model is then evaluated in the next quarter. Regret is the Test-2 net Sharpe of that quarter's best model minus that of the selected model. Under the fixed Top-10%, daily, 25-bp protocol, PPD lowers mean regret from 1.898 to 1.283 and mean rank from 4.00 to 3.05. Top-three retention rises from 40% to 65%.

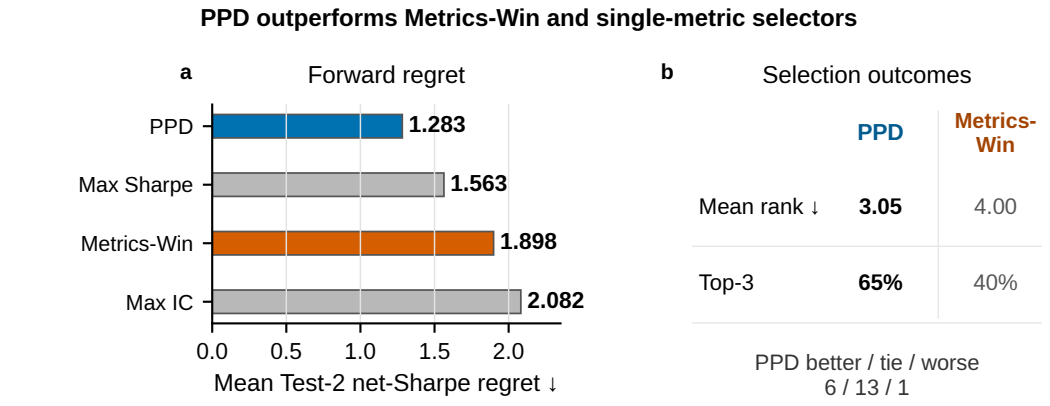

Figure 4: PPD versus Metrics-Win and two single-metric selectors under the fixed primary protocol (20 dependent pool-events). Max Sharpe and Max IC each select from one Test-1 metric. Panel b retains the direct PPD–Metrics-Win comparison; every outcome is evaluated only in the next quarter.

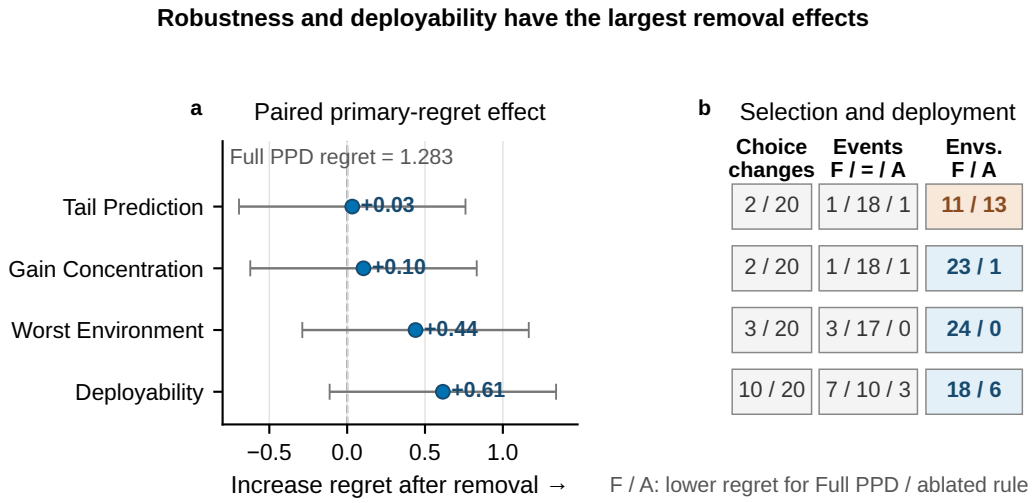

Figure 5: Leave-one-component-out diagnostic around the complete PPD. (a) Points are paired increases in mean Test-2 net-Sharpe regret after removal; whiskers are 95% quarter-cluster simultaneous intervals over the four contrasts. Positive values favor Full PPD. (b) Selection changes and descriptive lower/equal/higher-regret counts across 20 dependent pool-events and lower-regret counts across 24 correlated deployment environments.

Against Conventional Metrics-Win, PPD has lower, equal, and higher regret in 6, 13, and 1 events. Figure 5 removes four PPD-specific additions to the conventional Prediction and Portfolio core: tail-Prediction metrics, Gain Concentration, worst-environment Sharpe, and Deployability. The corresponding changes in primary regret are .033, .105, .439, and .615. Removing Deployability changes the selected model in 10 of 20 events, compared with 2–3 for the other components. Across the 24 deployment environments, complete PPD has lower regret than most ablated variants in most settings. Worst-environment robustness and Deployability thus have the largest effects on forward selection. Appendix Table 4 reports the factorial grid.

Answer to RQ2. PPD reduces primary regret from 1.898 to 1.283 and improves mean rank from 4.00 to 3.05. Among the four PPD-specific additions examined in the ablation, removing Deployability or worst-environment Portfolio evidence produces the largest effects on regret and model selection.

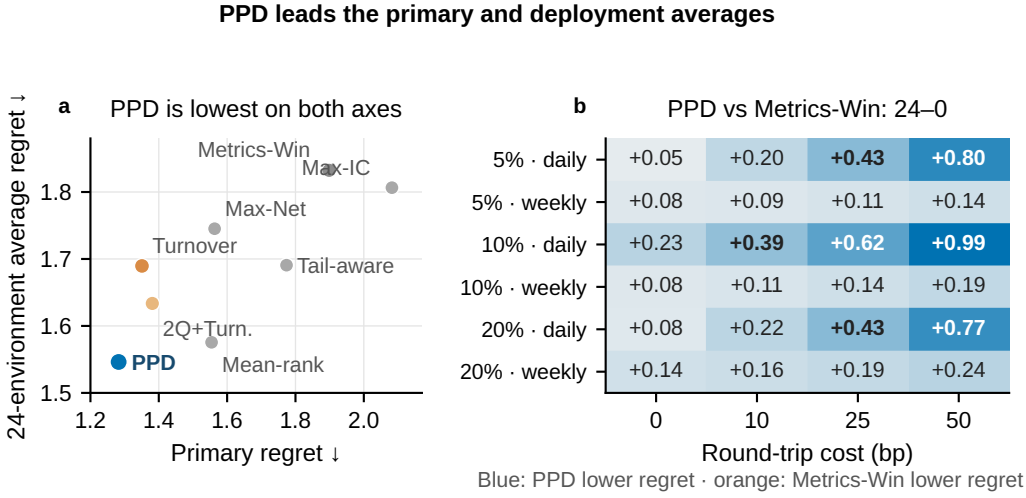

Figure 6: Test-2 deployment stress after one primary-protocol Test-1 selection. (a) Primary and 24-environment-average regret: PPD is lowest among eight rules on both axes. (b) Metrics-Win minus PPD regret across the $3 \times 2 \times 4$ holding-fraction, schedule, and cost grid; positive values favor PPD in all 24 cells.

5.3 RQ3: Does PPD remain useful under deployment shifts?

Allowing model reselection within each backtest setting would risk turning robustness evaluation into local backtest optimization. We therefore select each rule once under the primary Test-1 protocol and carry that choice unchanged into Test 2. Figure 6 evaluates that frozen choice under three holding fractions, two schedules, and four costs. PPD has lower mean regret than Conventional Metrics-Win in all 24 environments.

Answer to RQ3. PPD retains lower regret across the deployment grid; its lead over Metrics-Win holds in every setting and over the turnover-based controls in most settings. Together with the relatively large Deployability effect in the ablation, this suggests that turnover is informative but does not explain the result alone. PPD remains beneficial across the stress-test settings.

5.4 RQ4: What scientific superiority claims does the evidence support?

RQ3 evaluates whether the selected model remains useful across changing deployment conditions. RQ4 then applies a stricter standard for superiority claims by requiring the evidence to hold against the full declared comparator set under simultaneous inference. No model clears all six opponents under either the fixed or recalibrated full-family comparison, so the complete seven-model set has no qualified region. Figure 7 shows that a model can be selected for deployment without supporting a full-scope SOTA claim. Appendix Figure 8 reports the complete calibration audit.

Qualified cells contract as the comparator set expands. With four models, the fixed and recalibrated procedures identify 18 and 20 qualified pool–model cells, respectively. The counts fall to 12 and 14 when two models are added, then to zero with the seventh model. Recalibration therefore changes the amount of local support without changing the full-family conclusion. A model can clear a restricted comparator set while failing the conjunction required against all six opponents. This contraction may partly reflect the strength of the recent comparator pool. It also highlights that SOTA claims depend on baseline selection: superiority against a restricted set may disappear when additional strong competitors are included.

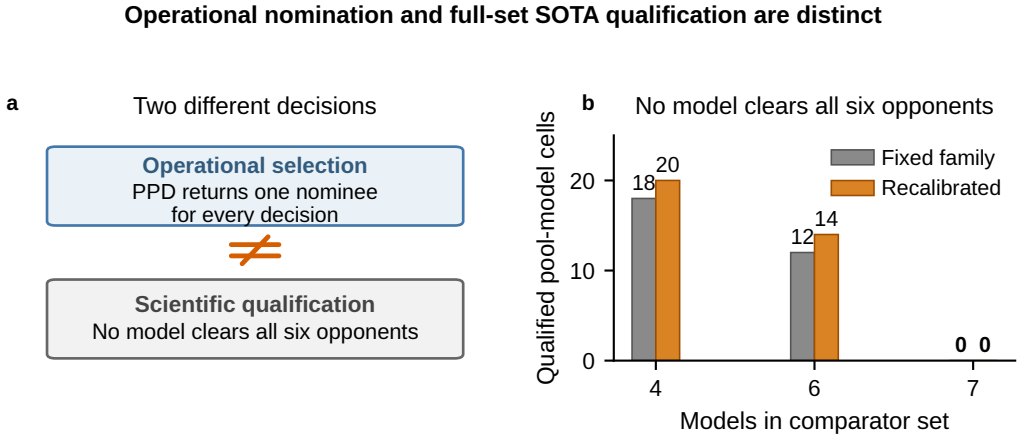

Figure 7: Deployment selection and full-set SOTA qualification answer different questions. (a) PPD provides a deployment nominee for the seven-model comparison. (b) Qualified pool–model cells shrink as the declared comparator set expands and reach zero for the complete seven-model set.

Restricted-scope superiority can still be reported when the comparator set, environment, practical margin, and interval family are specified. A broader SOTA claim requires the model to clear the corresponding expanded comparator family, rather than only a favorable pairwise comparison.

Answer to RQ4. None of the seven models clears all six declared opponents, so the full comparison does not support a full-scope SOTA claim. Superiority remains supported only within restricted comparator sets and environments.

6 Implications and Limitations

Practical model-selection implications. Conventional metrics-win rules are convenient, but they can be fragile in financial forecasting, where predictive signals must survive portfolio construction and changing out-of-sample trading conditions. PPD provides a more deployment-oriented basis for model selection by evaluating Prediction, Portfolio, and Deployability jointly. Rather than choosing a model from a single favorable metric or backtest setting, practitioners can assess whether its predictive advantage remains useful after portfolio translation and under realistic deployment conditions.

Research and reporting implications. For authors, PPD-SOTA encourages broader evaluation across prediction quality, portfolio performance, and deployability instead of optimizing toward a single headline metric or locally favorable backtest setting. For reviewers, the same structure provides a common basis for model comparison: forward performance informs deployment selection, while broader superiority claims are evaluated against a declared comparator set with uncertainty accounted for. This shifts the discussion away from isolated metric wins toward a more comparable standard for evaluating new financial forecasting models.

Limits. PPD-SOTA is not tied to a particular market: its core procedure only requires model scores, a portfolio construction rule, and a declared comparison set. The current empirical evidence, however, is concentrated on U.S. large-cap equities. The daily or weekly backtests also abstract from market depth, capacity, and endogenous price impact. PPD-SOTA can also be extended to broader agentic trading systems, although these settings require additional evaluation dimensions for agent-specific issues such as temporal information leakage and decision-process validity (Zhang et al., 2024a; Yu et al., 2024; Koa et al., 2024; Tang et al., 2025; Li et al., 2025c; Liu et al., 2026; Qin et al., 2024; Jeon et al., 2024; Zong et al., 2024; Pan et al., 2026).

7 Code and Data Availability

U.S. equity prices come from Financial Modeling Prep (FMP) and remain subject to its license. Original authors provide public source codes for all seven models. The paper and Appendix together provide the core specification of PPD-SOTA, including the model pool, account-replay procedure and parameters, selector components, and inference protocol. Upon acceptance, we will release the related scripts of PPD-SOTA evaluation.

8 Conclusion

This work revisits what a state-of-the-art evaluation should establish in financial AI. Existing benchmarks often use the same leaderboard to guide deployment and to support SOTA claims, although these decisions require different evidence. PPD-SOTA formalizes this separation through forward model selection, execution-consistent account replay, and complete-comparator qualification. Across seven recent forecasting models, PPD reduces next-quarter regret from 1.898 to 1.283 and improves deployment-average transfer over Metrics-Win. No model clears all six declared opponents. These results show that a useful operational choice and a defensible global SOTA claim need not coincide. The purpose is therefore not to replace one leaderboard with another, but to match each decision to the evidence it requires.

Beyond model ranking, PPD-SOTA provides an evaluation record. The PPD profile identifies whether weakness arises in Prediction, Portfolio, or Deployability, while scientific qualification records the comparator scope and uncertainty attached to a superiority claim. Future evaluations should test this protocol on broader universes, larger recent model families, and account replays that incorporate capacity and market impact. As financial AI benchmarks expand, the central distinction should remain: real-world deployment should be justified by forward operational evidence, whereas SOTA status should be limited to the comparison scope supported by statistical evidence.

AI USE STATEMENT

Generative AI tools were used to provide feedback on parts of the experimental design, assist with code implementation and debugging, and support interpretation of experimental results. All methodological choices, experimental protocols, analyses, and scientific claims were independently reviewed and finalized by the authors. The authors take full responsibility for the content of this work.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021.

David H. Bailey and Marcos López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014.

Vanya Priscillia Bendatu et al. Portfolio Reinforcement Learning with Scenario-Context Rollout. In *KDD*, 2026.

Zhiyuan Cao, Jiayu Xu, Chengqi Dong, Peiwen Yu, and Tian Bai. MATCC: A novel approach for robust stock price prediction incorporating market trends and cross-time correlations. In *ACM International Conference on Information and Knowledge Management*, pp. 187–196, 2024. doi: 10.1145/3627673.3679715.

Jared Chan et al. PortfolioPilot: An Agentic Platform for Financial Portfolio Management Algorithm Development and Evaluation. In *AAAI*, 2026.

Jiayi Chen et al. MARS: A Meta-Adaptive Reinforcement Learning Framework for Risk-Aware Multi-Agent Portfolio Management. In *AAAI*, 2026.

Weijun Chen et al. DHMoE: Diffusion Generated Hierarchical Multi-Granular Expertise for Stock Prediction. In *AAAI*, 2025.

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 *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.

Yitong Duan, Lei Wang, Qizhong Zhang, and Jian Li. FactorVAE: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. In *AAAI Conference on Artificial Intelligence*, 2022.

Yitong Duan, Weiran Wang, and Jian Li. FactorGCL: A hypergraph-based factor model with temporal residual contrastive learning for stock returns prediction. In *AAAI Conference on Artificial Intelligence*, 2025.

Jinyong Fan and Yanyan Shen. StockMixer: A simple yet strong MLP-based architecture for stock price forecasting. In *AAAI Conference on Artificial Intelligence*, pp. 8389–8397, 2024. doi: 10.1609/aaai.v38i8.28681.

Siyu Gao et al. StockFormer: Learning Hybrid Trading Machines with Predictive Coding. In *IJCAI*, 2023.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79 (2):453–497, 2011.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Yoontae Hwang et al. Signature-Informed Transformer for Asset Allocation. In *ICML*, 2026.

Hinal Jajal et al. Global Merger-Arbitrage Forecasting with Language Models. In *ICML*, 2026.

Jihyeong Jeon, Jiwon Park, Chanhee Park, and U Kang. FreQuant: A reinforcement-learning based adaptive portfolio optimization with multi-frequency decomposition. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. doi: 10.1145/3637528.3671668.

Kelvin J. L. Koa, Ritchie Ng, Yunshan Ma, and Tat-Seng Chua. Learning to generate explainable stock predictions using self-reflective large language models. In *The Web Conference*, 2024. doi: 10.1145/3589334.3645611.

Yaxuan Kong et al. Evaluating LLMs in Finance Requires Explicit Bias Consideration. In *ICML*, 2026.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Changlun Li, Yao Shi, Chen Wang, Qiqi Duan, Runke Ruan, Weijie Huang, Haonan Long, Lijun Huang, Nan Tang, and Yuyu Luo. Time travel is cheating: Going live with DeepFund for real-time fund investment benchmarking. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025a.

Shicheng Li, Jinshan Zhang, and Feng Wang. AlphaGAT: A two-stage learning approach for adaptive portfolio selection. In *International Joint Conference on Artificial Intelligence*, 2025b.

Shuqi Li et al. PEN: Prediction-Explanation Network to Forecast Stock Price Movement with Better Explainability. In *AAAI*, 2023.

Shuqi Li et al. CausalStock: Deep End-to-end Causal Discovery for News-driven Stock Movement Prediction. In *NeurIPS*, 2024a.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-guided stock transformer for stock price forecasting. In *AAAI Conference on Artificial Intelligence*, 2024b.

Weixian Waylon Li et al. Can LLM-based Financial Investing Strategies Outperform the Market in Long Run? In *KDD*, 2026.

Yuante Li, Xu Yang, Xiao Yang, Minrui Xu, 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*, 2025c.

Zhengze Li, Zilin Song, Tingting Yuan, and Xiaoming Fu. COGRASP: Co-occurrence graph based stock price forecasting. In *International Joint Conference on Artificial Intelligence*, 2025d.

Zhiming Li et al. Logic-Q: Improving Deep Reinforcement Learning-based Quantitative Trading via Program Sketch-based Tuning. In *AAAI*, 2025e.

Yun-Hsuan Lien et al. Contrastive Learning and Reward Smoothing for Deep Portfolio Management. In *IJCAI*, 2023.

Thomas Y. L. Lin et al. Latent Variable Estimation in Bayesian Black-Litterman Models. In *ICML*, 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 *Annual Meeting of the Association for Computational Linguistics*, 2026.

Mengpu Liu et al. ECHO-GL: Earnings Calls-Driven Heterogeneous Graph Learning for Stock Movement Prediction. In *AAAI*, 2024.

Xiao-Yang Liu, Ziyi Xia, Hongyang Yang, Jiechao Gao, Daochen Wang, Li Zha, Ming Guo, Xiaomin Xiong, and Jianyang Wang. FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Di Luo et al. Causality-Guided Multi-Memory Interaction Network for Multivariate Stock Price Movement Prediction. In *ACL*, 2023.

Simon Mahns et al. Joint-Embedding Predictive Learning of Latent Market States in U.S. Equities. In *ICML*, 2026.

Jieyuan Mei, Jindong Tian, Ronghui Xu, Hanyue Wei, Chenjuan Guo, and Bin Yang. FinD3: A dual 3d state space model with dynamic hypergraph for financial stock prediction. In *ACM International Conference on Information and Knowledge Management*, 2025. doi: 10.1145/3746252.3761239.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. Training on the benchmark is not all you need. In *AAAI Conference on Artificial Intelligence*, 2025.

Chaofan Pan, Lingfei Ren, Linbo Xiong, Yonghao Li, Wei Wei, and Xin Yang. Regime-adaptive continual learning for portfolio management. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3817620.

Feiyang Pan et al. Learn Continuously, Act Discretely: Hybrid Action-Space Reinforcement Learning For Optimal Execution. In *IJCAI*, 2022.

Kijeong Park et al. Dynamic Higher-Order Relations and Event-Driven Temporal Modeling for Stock Price Forecasting. In *IJCAI*, 2025.

Zhiyuan Peng et al. FNSPID: A Comprehensive Financial News Dataset in Time Series. In *KDD*, 2024.

Dimitris N. Politis and Joseph P. Romano. A circular block-resampling procedure for stationary data. In *Exploring the Limits of Bootstrap*, pp. 263–270. Wiley, 1992.

Molei Qin, Shuo Sun, Wentao Zhang, Haochong Xia, Xinrun Wang, and Bo An. EarnHFT: Efficient hierarchical reinforcement learning for high frequency trading. In *AAAI Conference on Artificial Intelligence*, pp. 14669–14676, 2024.

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 *AAAI Conference on Artificial Intelligence*, 2025.

Yu Shi, Zongliang Fu, Shuo Chen, Bohan Zhao, Wei Xu, Changshui Zhang, and Jian Li. Kronos: A foundation model for the language of financial markets. In *AAAI Conference on Artificial Intelligence*, 2026.

Chen-Hui Song, Shuoling Liu, and Liyuan Chen. The label horizon paradox: Rethinking supervision targets in financial forecasting. In *International Conference on Machine Learning*, 2026.

Shuo Sun, Molei Qin, Wentao Zhang, Haochong Xia, Chuqiao Zong, Jie Ying, Yonggang Xie, Lingxuan Zhao, Xinrun Wang, and Bo An. TradeMaster: A holistic quantitative trading platform empowered by reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023a.

Shuo Sun et al. Mastering Stock Markets with Efficient Mixture of Diversified Trading Experts. In *KDD*, 2023b.

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 *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. doi: 10.1145/3711896.3736838.

Heyuan Wang et al. Adaptive Long-Short Pattern Transformer for Stock Investment Selection. In *IJCAI*, 2022a.

Heyuan Wang et al. Heterogeneous Interactive Snapshot Network for Review-Enhanced Stock Profiling and Recommendation. In *IJCAI*, 2022b.

Saizhuo Wang, Hao Kong, Jiadong Guo, Fengrui Hua, Yiyan Qi, Wanyun Zhou, Jiahao Zheng, Xinyu Wang, Lionel M. Ni, and Jian Guo. QuantBench: Benchmarking AI methods for quantitative investment. *arXiv preprint arXiv:2504.18600*, 2025.

Zikai Wei et al. HireVAE: An Online and Adaptive Factor Model Based on Hierarchical and Regime-Switch VAE. In *IJCAI*, 2023.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

Haochong Xia et al. Bayesian Robust Financial Trading with Adversarial Synthetic Market Data. In *KDD*, 2026.

Hongjie Xia, Huijie Ao, Long Li, Yu Liu, Sen Liu, Guangnan Ye, and Hongfeng Chai. CI-STHPAN: Pre-trained attention network for stock selection with channel-independent spatio-temporal hypergraph. In *AAAI Conference on Artificial Intelligence*, pp. 9187–9195, 2024. doi: 10.1609/aaai.v38i8.28770.

Chen Yang, Jingyuan Wang, Xiaohan Jiang, and Junjie Wu. Learning universal multi-level market irrationality factors to improve stock return forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. doi: 10.1145/3690624.3709328.

Linyi Yang et al. NumHTML: Numeric-Oriented Hierarchical Transformer Model for Multi-Task Financial Forecasting. In *AAAI*, 2022a.

Linyi Yang et al. Measuring Consistency in Text-based Financial Forecasting Models. In *ACL*, 2023a.

Mengyuan Yang, Xiaolin Zheng, Qianqiao Liang, Bing Han, and Mengying Zhu. A smart trader for portfolio management based on normalizing flows. In *International Joint Conference on Artificial Intelligence*, 2022b.

Mengyuan Yang et al. Spotlight News Driven Quantitative Trading Based on Trajectory Optimization. In *IJCAI*, 2023b.

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 *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5476–5486, 2023. doi: 10.1145/3580305.3599831.

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 *Advances in Neural Information Processing Systems*, 2024.

Haochen Yuan et al. MetaTrader: Learning to Generalize RL Trading Policies Beyond Offline Data. In *AAAI*, 2026.

Chuheng Zhang, Yitong Duan, Xiaoyu Chen, Jianyu Chen, Jian Li, and Li Zhao. Towards generalizable reinforcement learning for trade execution. In *International Joint Conference on Artificial Intelligence*, 2023.

Julian Zhong-Nan Zhang et al. An Interactive Simulation Framework by Ensemble Imitation Learning Agents for Training Robust Trading Policies. In *AAAI*, 2026a.

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 *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024a. doi: 10.1145/3637528.3671801.

Wentao Zhang, Yilei Zhao, Chuqiao Zong, Xinrun Wang, and Bo An. FinWorld: An all-in-one open-source platform for end-to-end financial AI research and deployment. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026b. doi: 10.1145/3770854.3785689.

Wentao Zhang et al. Reinforcement Learning with Maskable Stock Representation for Portfolio Management in Customizable Stock Pools. In *WWW*, 2024b.

Xu Zhang et al. Multi-period Learning for Financial Time Series Forecasting. In *KDD*, 2025a.

Zexi Zhang, Lifan Zhao, Yanyan Shen, and Bin Yao. StockCL: Selective contrastive learning for stock trend forecasting via learnable concepts. In *Database Systems for Advanced Applications*, pp. 313–329, 2024c. doi: 10.1007/978-981-97-5575-2_20.

Ziliang Zhang et al. Online Portfolio Selection with ML Predictions. In *NeurIPS*, 2025b.

Lifan Zhao, Shuming Kong, and Yanyan Shen. DoubleAdapt: A meta-learning approach to incremental learning for stock trend forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3492–3503, 2023. doi: 10.1145/3580305.3599315.

Zhiyuan Zhao et al. Performative Time-Series Forecasting. In *KDD*, 2025.

Xiaolin Zheng et al. Deep Hashing-based Dynamic Stock Correlation Estimation via Normalizing Flow. In *IJCAI*, 2023.

Haoren Zhu et al. Forecasting Asset Dependencies to Reduce Portfolio Risk. In *AAAI*, 2022.

Chuqiao Zong, Chaojie Wang, Molei Qin, Lei Feng, Xinrun Wang, and Bo An. MacroHFT: Memory augmented context-aware reinforcement learning on high frequency trading. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. doi: 10.1145/3637528.3672064.

Chuqiao Zong, Molei Qin, Haochong Xia, and Bo An. ArchetypeTrader: Reinforcement learning for selecting and refining learnable strategic archetypes in quantitative trading. In *AAAI Conference on Artificial Intelligence*, 2026.

A EXACT EVIDENCE IDENTITY

The analysis uses 28 frozen prediction paths and 336 complete account cells. The manifest records candidate, pool, seed, split, source, prediction result, account policy, cost, and statistic code.

Table 2: Fixed same-task model pool. All methods have author-source implementations and publication-level performance evidence.

Model	Year	Venue	Preserved mechanism
MASTER	2024	AAAI	market-guided cross-stock Transformer
StockMixer	2024	AAAI	indicator/time/stock MLP mixing
CI-STHPAN-DTW	2024	AAAI	pretraining and dynamic hypergraph
MATCC	2024	CIKM	market trend and cross-time correlation
StockCL-Transformer	2024	DASFAA	concept-selective contrastive learning
UMI	2025	KDD	multi-level market irrationality factors
FinD3-DTW	2025	CIKM	dual state-space model and dynamic hypergraph

Computational interpretation. The 28 prediction paths are the costly trained units. The deployment grid consists of cached portfolio replays, and the 10,000 bootstrap repetitions reuse shared return arrays.

Minimum reporting record. A decision record includes candidate/source identifiers, prediction and exclusion results, Test-1 cutoff and Test-2 dates, label and portfolio timing, costs, missing-data policy, metric definitions, layer weights, tie rule, and comparison family. The scientific comparison also records its margin, block length, confidence, bootstrap seed, interval family, and four-state cells.

B ACCOUNT REPLAY AND FROZEN PARAMETERS

Algorithm 1 Causal top-fraction account replay with proportional costs

Require: two prediction paths; membership; adjusted opens; fraction q; frequency f; round-trip cost b

Start with one dollar in cash and set $c \leftarrow b/20000$
for each signal day t **do**
  Rank each seed across stocks and average the two percentile ranks
  if open $t+1$ is a trading day under f **then**
    Select the top $k = \lceil qN_t \rceil$ eligible stocks; set each target to $1/k$
    Value existing holdings; fail if a held stock has no defensible mark
    Cancel any unquoted new purchase and leave its intended value in cash
    Solve $A + c\sum_i |Aw_i - v_i| = W^-$ and trade only $Aw_i - v_i$
  else
    Keep shares unchanged and charge no trading fee
  end if
  Record gross/net return, fees, traded notional, cash, and exposure
end for
Liquidate remaining positions at the endpoint and charge the final cost

Table 3: Frozen account-replay parameters. Turnover is gross absolute buy-plus-sell notional divided by pre-trade equity.

Item	Frozen setting
Universes	NASDAQ-100 and DJ30; signal-date membership
Forecasts	Seven models; seeds 101/211 fused by mean daily percentile rank
Timing	Score after close t; trade at open $t+1$; target open$[t+2]$/open$[t+1]-1$
Portfolio	Long-only top 5/10/20%; equal target weights
Rebalancing / costs	Daily or weekly; 0/10/25/50-bp round trip
Missing quote	New target cancelled to cash; missing held mark makes path unavailable
Boundary / ties	Final two signals purged; endpoint liquidation; frozen-ID tie order

C SELECTOR DIAGNOSTICS

The complete factorial grid compares component combinations with the matched equal-weight control. The metric set, layer assignments, and 40/30/30 weights were fixed before Test-2 outcomes were examined. The subsequent factorial and weight grids assess sensitivity under the reported specification.

Table 4: Complete 2^4 component sensitivity grid plus the matched equal-weight full-metric control. T: tail Prediction; C: inverse Gain Concentration; W: Test-1 worst-environment Sharpe; D: Deployability. The factorial rows retain the frozen 40/30/30 layer rule, renormalized to 4/7 and 3/7 when D is absent. Lower regret is better; rows are ordered by configuration, not outcome.

Configuration	T	C	W	D	Primary	Deploy. avg.	Worst env.
Core	–	–	–	–	1.605	1.583	1.954
D	–	–	–	✓	1.694	1.824	2.701
W	–	–	✓	–	1.981	1.898	2.307
C	–	✓	–	–	1.949	1.783	2.327
T	✓	–	–	–	1.818	1.789	2.325
W+D	–	–	✓	✓	1.381	1.634	2.566
C+D	–	✓	–	✓	1.315	1.537	2.505
C+W	–	✓	✓	–	2.100	1.949	2.346
T+D	✓	–	–	✓	1.751	1.836	2.638
T+W	✓	–	✓	–	1.878	1.741	2.299
T+C	✓	✓	–	–	1.842	1.814	2.342
C+W+D	–	✓	✓	✓	1.315	1.537	2.505
T+W+D	✓	–	✓	✓	1.387	1.610	2.503
T+C+D	✓	✓	–	✓	1.722	1.785	2.663
T+C+W	✓	✓	✓	–	1.897	1.778	2.179
Full PPD	✓	✓	✓	✓	**1.283**	**1.546**	**2.431**
Equal-weight full	✓	✓	✓	✓	1.466	1.667	2.633

With the layer definitions and selection events fixed, most tested weights yield lower mean next-quarter regret than Metrics-Win. The reported weights were fixed before the sensitivity sweep; the sweep optimum occurs at a different setting.

D INFERENCE PROTOCOL AND CALIBRATION

Primary inference uses circular 20-session blocks, 10,000 bootstrap repetitions, 95% simultaneous confidence, and $\delta = .10$. Sensitivity covers δ of .05/.10/.20, block lengths 5/20/60, confidence 90/95/99%, and four windows.

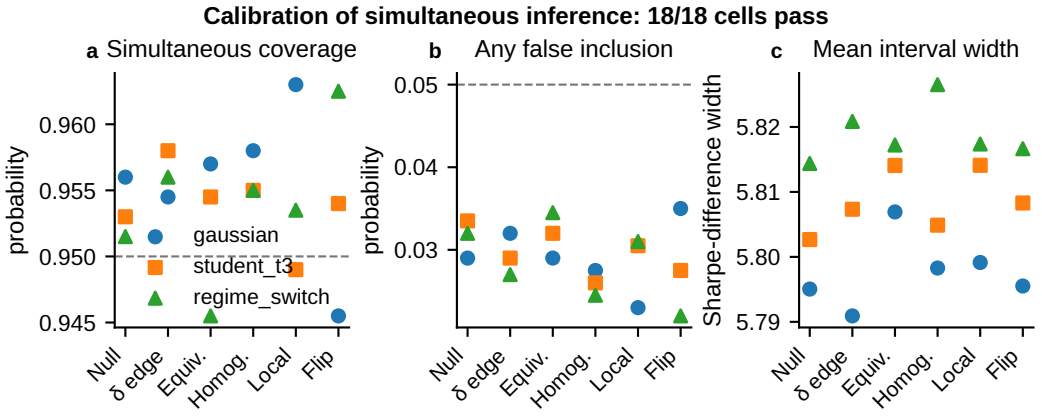

Figure 8: Calibration of simultaneous inference. Production inference is applied to joint return paths; dashed lines mark .95 simultaneous coverage and .05 false inclusion.

E SIXTY-FIVE-PAPER METRIC AND SOTA-STANDARD AUDIT

The corpus samples 65 financial-AI papers published in 2022–2026 at seven leading computer-science conferences: AAAI, ACL, ICML, IJCAI, KDD, NeurIPS, and WWW. Table 5 organizes the frozen literature scan by the distinction used in this paper: prediction metrics, the mapping from signals to a backtest, portfolio outcomes, and the standard used to state superiority.

Table 5: Metrics, backtest strategies, and claimed SOTA standards in the 65-paper audit. Prediction and portfolio metrics are separated from each paper's recovered metric list. "Metric rank" records the form of the automatically recovered superiority claim; it is not an endorsement that the claim satisfies PPD-SOTA.

Ref.	Method	Prediction metrics	Backtest strategy	Portfolio metrics	SOTA standard
(Zhu et al., 2022)	Asset Dependencies	Accuracy, MSE	Top-k	Volatility	Metric rank: baseline win (1)
(Duan et al., 2022)	FactorVAE	IC, RankIC, ICIR	Top-k	Return, PnL, Sharpe, MDD, Turnover	Metric rank: baseline win (2)
(Yang et al., 2022a)	NumHTML	Accuracy, MCC, MSE	UNR	PnL, Sharpe, Volatility	Metric rank: explicit SOTA (1)
(Wang et al., 2022a)	ALST	UNR	Long–short	Return, PnL, Sharpe, Volatility	No superiority claim
(Pan et al., 2022)	Hybrid Execution	NA	Execution policy	PnL	Metric rank: baseline win (1)
(Yang et al., 2022b)	Smart Trader	NA	Multiple	PnL, Sharpe, MDD, Volatility, Calmar	Metric rank: baseline win (2)
(Wang et al., 2022b)	HISN	UNR	UNR	Return, PnL, Sharpe	Metric rank: baseline win (1)
(Liu et al., 2022)	FinRL-Meta	NA	Buy-and-hold	Return, PnL, Sharpe, MDD, Volatility, Calmar	No superiority claim
(Li et al., 2023)	PEN	Accuracy, MCC	Top-k	PnL, Volatility	Metric rank: baseline win (3)
(Luo et al., 2023)	CMIN	Accuracy, MCC	NA	NA	Metric rank: explicit SOTA (2)
(Yang et al., 2023a)	FinTrust	Accuracy, MCC	NA	PnL, Sharpe, Volatility	No superiority claim
(Lien et al., 2023)	CPM	NA	UNR	PnL, Sharpe, MDD, Volatility	No superiority claim
(Zheng et al., 2023)	HDCF	Accuracy	UNR	Return, MDD, Volatility	Metric rank: baseline win (2)
(Zhang et al., 2023)	Generalizable TE	NA	Execution policy	Volatility	Metric rank: baseline win (1)
(Wei et al., 2023)	HireVAE	IC, RankIC, ICIR	UNR	UNR	Metric rank: explicit SOTA (1)
(Yang et al., 2023b)	Spotlight Trader	NA	Top-k	Return, PnL, Sharpe, MDD, Sortino	Metric rank: baseline win (1)
(Gao et al., 2023)	StockFormer	NA	Buy-and-hold	PnL, Sharpe, MDD	Metric rank: baseline win (4)
(Sun et al., 2023b)	AlphaMix	Accuracy	Top-k	PnL, Sharpe, MDD, Volatility, Sortino	Metric rank: explicit SOTA (3)
(Zhao et al., 2023)	DoubleAdapt	IC, RankIC, ICIR, Accuracy	UNR	Return, PnL	Metric rank: baseline win (1)
(Sun et al., 2023a)	TradeMaster	NA	NA	PnL, Sharpe, MDD, Volatility, Calmar, Sortino	No superiority claim
(Qin et al., 2024)	EarnHFT	NA	Execution policy	PnL, Sharpe, MDD, Volatility, Calmar, Sortino	Metric rank: baseline win (2)

Continued

Table 5 continued.

Ref.	Method	Prediction metrics	Backtest strategy	Portfolio metrics	SOTA standard
(Liu et al., 2024)	ECHO-GL	Accuracy, MCC	NA	PnL, Sharpe	No superiority claim
(Li et al., 2024b)	MASTER	IC, RankIC, ICIR, Accuracy	Top-k	Return, PnL, Volatility	Metric rank: baseline win (1)
(Fan & Shen, 2024)	StockMixer	IC, Accuracy	Top-k	PnL, Sharpe, Volatility	Metric rank: baseline win (1)
(Zhang et al., 2024a)	FinAgent	Accuracy	Multiple	Return, PnL, Sharpe, MDD, Volatility, Calmar, Sortino	Metric rank: explicit SOTA (1)
(Peng et al., 2024)	FNSPID	Accuracy, MSE, MAE	Multiple	PnL, Sharpe, Volatility	No superiority claim
(Jeon et al., 2024)	FreQuant	NA	Multiple	PnL, Sharpe, Volatility	Metric rank: baseline win (2)
(Zong et al., 2024)	MacroHFT	NA	Multiple	PnL, Sharpe, MDD, Volatility, Calmar	Metric rank: explicit SOTA (2)
(Li et al., 2024a)	CausalStock	Accuracy, MCC	Top-k	UNR	Metric rank: baseline win (2)
(Yu et al., 2024)	FINCON	NA	Buy-and-hold	Return, PnL, Sharpe, MDD	Metric rank: baseline win (1)
(Zhang et al., 2024b)	MSR	UNR	Multiple	Return, PnL, MDD, Volatility	Metric rank: baseline win (2)
(Koa et al., 2024)	Self-Reflective LLM	Accuracy, MCC	Top-k	UNR	Metric rank: baseline win (2)
(Shi et al., 2025)	AlphaForge	IC, RankIC	UNR	UNR	Metric rank: baseline win (1)
(Chen et al., 2025)	DHMoE	UNR	UNR	Return, PnL, Sharpe	Metric rank: explicit SOTA (1)
(Duan et al., 2025)	FactorGCL	IC, ICIR	UNR	PnL	Metric rank: explicit SOTA (2)
(Li et al., 2025e)	Logic-Q	NA	Multiple	Return, Sharpe, MDD, Volatility	No superiority claim
(Ni et al., 2025)	TOTB	Accuracy	NA	NA	No superiority claim
(Lin et al., 2025)	Bayesian BL	NA	Direct weights	Return, Sharpe, MDD, Turnover, Volatility	No superiority claim
(Li et al., 2025b)	AlphaGAT	IC, RankIC, Accuracy	Top-k	PnL	Metric rank: explicit SOTA (1)
(Li et al., 2025d)	COGRASP	IC, RankIC, ICIR, Accuracy	Top-k	Turnover	Metric rank: explicit SOTA (1)
(Park et al., 2025)	DHOR	IC, RankIC, Accuracy	Top-k	PnL, Sharpe, MDD	Metric rank: baseline win (1)
(Tang et al., 2025)	AlphaAgent	IC, RankIC, ICIR	Top-k	Return, PnL, MDD	Metric rank: baseline win (1)
(Zhang et al., 2025a)	MPL	Accuracy, MSE, MAE	NA	PnL, Sharpe	Metric rank: baseline win (4)
(Zhao et al., 2025)	Performative TS	Accuracy, MSE, MAE	NA	NA	Metric rank: baseline win (3)
(Li et al., 2025a)	DeepFund	NA	NA	Return, PnL, Sharpe, MDD	No superiority claim
(Zhang et al., 2025b)	OPS-ML	Accuracy	UNR	Volatility, Regret	No superiority claim
(Li et al., 2025c)	R&D-Agent-Quant	IC, RankIC, ICIR, Accuracy	Long–short	Return, MDD, Volatility, Calmar	Metric rank: explicit SOTA (1)
(Zong et al., 2026)	ArchetypeTrader	NA	UNR	PnL, Sharpe, MDD, Volatility, Calmar, Sortino, Regret	Metric rank: explicit SOTA (3)
(Zhang et al., 2026a)	Ensemble IL Sim	MSE	Buy-and-hold	PnL, Sharpe, MDD, Turnover, Volatility	Metric rank: baseline win (2)

Continued

Table 5 continued.

Ref.	Method	Prediction metrics	Backtest strategy	Portfolio metrics	SOTA standard
(Shi et al., 2026)	Kronos	IC, RankIC, Accuracy, MSE, MAE	Top-k	Volatility	Metric rank: baseline win (1)
(Chen et al., 2026)	MARS	NA	UNR	Return, PnL, Sharpe, MDD, Volatility	No superiority claim
(Yuan et al., 2026)	MetaTrader	NA	Buy-and-hold	Return, Sharpe, MDD	Metric rank: baseline win (1)
(Chan et al., 2026)	PortfolioPilot	MSE, MAE	UNR	Sharpe, MDD, Turnover	No superiority claim
(Liu et al., 2026)	CAM	IC, RankIC, ICIR, Accuracy	UNR	Return, Volatility	No superiority claim
(Kong et al., 2026)	Finance LLM Bias	Accuracy	NA	PnL	No superiority claim
(Mahns et al., 2026)	Market JEPA	NA	NA	Volatility	No superiority claim
(Song et al., 2026)	Label Horizon	IC, RankIC, ICIR, Accuracy	UNR	Return, Sharpe, MDD	Metric rank: explicit SOTA (1)
(Jajal et al., 2026)	Merger-Arb LM	Accuracy, MSE	NA	PnL	No superiority claim
(Hwang et al., 2026)	SIT	Accuracy	Direct weights	Sharpe, MDD, Volatility, Sortino	Metric rank: baseline win (2)
(Ding et al., 2026)	AlphaEval	IC, RankIC, ICIR, Accuracy	Top-k	Return, Sharpe, MDD, Turnover, Volatility	No superiority claim
(Xia et al., 2026)	Bayesian Robust	NA	Buy-and-hold	Return, PnL, MDD, Volatility	Metric rank: explicit SOTA (1)
(Li et al., 2026)	FinSABER	NA	Buy-and-hold	Return, PnL, MDD, Volatility	No superiority claim
(Zhang et al., 2026b)	FinWorld	RankIC, Accuracy, MSE, MAE	Top-k	Return, PnL, Sharpe, MDD, Volatility	No superiority claim
(Pan et al., 2026)	RACL	NA	Multiple	Return, PnL, Sharpe, MDD, Volatility	Metric rank: explicit SOTA (1)
(Bendatu et al., 2026)	SC-RL	NA	Direct weights	Return, Sharpe, MDD, Turnover, Volatility, Calmar	No superiority claim

Coding note. UNR means the field may apply but was not safely recovered from the source; NA means the field does not apply to that study. Counts in parentheses are automatically recovered claim counts, not independent statistical tests.