Forecast-to-Trade: Hierarchical Reinforcement Learning for Decision-Aware Financial Forecasting

Zijie Zhao [1] **Roy E. Welsch** [1]

Abstract

Forecasting systems are often judged by prediction quality, but real deployments require converting forecasts into actions under constraints, costs, and risk. We study this forecast-to-decision problem in text-aware equity portfolio management, where market forecasts and news-derived risk signals must become executable rebalancing decisions. We propose **Hierarchical Reinforced Trader (HRT)**, a bi-level reinforcement learning framework that separates sparse directional selection from risk-aware portfolio execution. A factorized High-Level Controller selects per-asset increase, reduce, or hold directions from compact market and text-derived signals, while a Low-Level Controller converts these directions into feasible portfolio weight adjustments under turnover, drawdown, and text-risk penalties. On an open 89-stock Nasdaq news benchmark with 2013–2018 training, 2019 validation, and 2020–2023 out-of-sample testing, HRT improves the learning-based return–risk–cost trade-off: Sharpe rises from 1.06 for HRT-Base to 1.24, daily turnover falls from 0.112 to 0.090, and the policy remains comparatively robust under transaction-cost stress. These results position financial trading as a demanding benchmark for decision-aware AI forecasting rather than point prediction alone.

1. Introduction

Forecasting is useful only when it can inform decisions. In practice, decision makers ask not only whether a forecast is accurate, but how an uncertain prediction should be acted on under budgets, frictions, risk limits, and delayed feedback. Quantitative finance is a natural testbed: time-series forecasts, news signals, and model-based risk estimates are available, yet portfolio performance depends on how signals are converted into trades after costs and downside risk.

This paper studies decision-aware financial forecasting in text-aware multi-asset portfolio control. Classical portfolio theory formalizes the return–risk trade-off (Markowitz, 1952), and reinforcement learning (RL) provides a sequential framework for trading under changing states (Bertsekas, 2012; Yang et al., 2020a; Liu et al., 2021). Recent financial news datasets and language-model pipelines provide sentiment, event, and risk-related signals from news (Zhang et al., 2023; Dong et al., 2024; Benhenda, 2025). Yet a high-confidence return forecast or useful text-risk signal does not automatically define a good action. A policy must decide which assets to act on, how much to trade, and when the expected benefit is worth the turnover, drawdown, and execution cost.

Flat trading policies blur these responsibilities. A single policy mapping all signals to portfolio weights must learn selection, sizing, cost control, and risk management at once. In large universes, the joint action space grows quickly, inputs are noisy, and reacting too aggressively to forecasts can erase alpha through turnover. We instead treat trading as a structured decision layer around forecasts.

We propose **Hierarchical Reinforced Trader (HRT)**, a forecast-conditioned hierarchical RL framework for text-aware portfolio management. The High-Level Controller (HLC) forms sparse directions—increase, reduce, or hold for each asset. The Low-Level Controller (LLC) translates those directions into feasible portfolio weight changes under cost, turnover, drawdown, and text-risk considerations. This yields an inspectable two-stage forecast-to-action pipeline without enumerating the full joint direction space.

Our contributions are threefold. First, we frame text-aware portfolio management as decision-aware forecasting, where upstream market and news-derived forecasts are judged through realized portfolio decisions under frictions. Second, we introduce a bi-level RL architecture that factorizes sparse directional selection and risk-aware continuous execution. Third, we evaluate the method on an open stock-news benchmark with same-universe baselines, ablations, and cost stress tests, showing improved return–risk–cost

[1]Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Zijie Zhao <zijiezha@mit.edu>, Roy E. Welsch <rwelsch@mit.edu>.

Accepted for spotlight presentation at the ICML 2026 Workshop on Forecasting as a New Frontier of Intelligence.

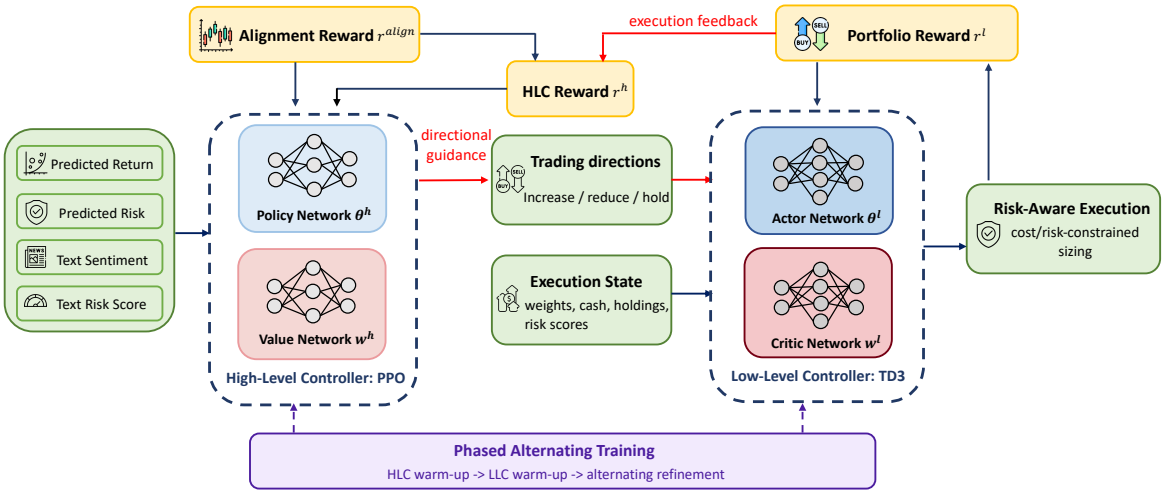

Figure 1. Overview of HRT as a forecast-to-decision system. Structured market forecasts and benchmark-provided text signals enter a factorized high-level controller that selects sparse trading directions. A low-level controller converts these directions into feasible, cost- and risk-aware portfolio adjustments, allowing execution feedback to shape future directional decisions.

trade-offs among learning-based strategies without claiming dominance on every risk metric.

2. Forecast-Conditioned Hierarchical Control

Structured signal interface. HRT is agnostic to the upstream forecasting model. For asset i on trading day t, the high-level state is

$$s^h_{i,t} = [\bar{w}_{i,t}, \hat{r}_{i,t}, \hat{\sigma}_{i,t}, u_{i,t}, \rho_{i,t}], \tag{1}$$

where $\bar{w}_{i,t}$ is the current pre-trade weight, $\hat{r}_{i,t}$ a predicted forward return, $\hat{\sigma}_{i,t}$ a realized-volatility proxy, $u_{i,t}$ a structured textual sentiment score, and $\rho_{i,t}$ a structured textual risk score. In our experiments, $\hat{r}_{i,t}$ comes from a LightGBM model trained on OHLCV-derived technical indicators and Qlib-style factors (Ke et al., 2017; Yang et al., 2020b); $\hat{\sigma}_{i,t}$ is the 20-day rolling realized volatility using past returns only. Text sentiment and risk are benchmark-provided signals from timestamp-aligned financial news, normalized using training-period statistics and cached before RL training. No LLM fine-tuning or online LLM querying is performed during training or evaluation.

For each decision, the policy observes only market and text-derived signals available before the close-of-day cutoff on day t. The portfolio is rebalanced at the next tradable open; realized returns are used only as ex-post rewards or evaluation targets. After-cutoff or non-trading-day news is rolled forward to the next tradable decision date, preserving HRT as a forecast-conditioned policy rather than a hindsight labeler.

Factorized high-level controller. The HLC selects stock-level trading directions $a^h_{i,t} \in \{-1, 0, 1\}$, denoting reduce, hold, and increase. Instead of treating the daily direction vector as one unstructured action over 3^N possibilities, HRT uses a factorized categorical policy

$$\pi^h_\theta(a^h_t \mid s^h_t) = \prod_{i=1}^{N} \pi^h_\theta(a^h_{i,t} \mid s^h_{i,t}). \tag{2}$$

The policy uses shared parameters and asset-level logits, evaluating $O(N)$ categorical distributions rather than a joint 3^N action space. Sparse activation comes from the hold action, a validation-selected confidence filter, and an activation penalty. The high-level reward combines directional alignment with downstream execution feedback,

$$r^h_t = \alpha_t \tilde{r}^{align}_t + (1 - \alpha_t)\tilde{r}^l_t - \lambda_{act} \frac{|\mathcal{A}^{exec}_t|}{N}, \tag{3}$$

where $\mathcal{A}^{exec}_t$ is the set of assets with nonzero executed weight changes. Alignment is computed only over executed changes. Next-day realized returns are used for ex-post rewards and metrics, but never enter the observed state at decision time.

Risk-aware low-level controller. Given the HLC directions, the LLC determines weight adjustments. Let $\bar{w}_t$ denote pre-trade weights, b_t cash, V_t portfolio value, and $d_t = (V_t^{peak} - V_t)/V_t^{peak}$ the drawdown state. The LLC observes

$$s_t^l = [\bar{w}_t, b_t/V_t, d_t, a_t^h, \hat{r}_t, \hat{\sigma}_t, \rho_t]. \tag{4}$$

A TD3-style actor first proposes a raw adjustment $\widetilde{\Delta w}_t = \mu_\phi(s_t^l)$ (Fujimoto et al., 2018). A deterministic feasibility layer then maps it to the executed adjustment,

$$\Delta w_t^{exec} = \mathcal{F}(\widetilde{\Delta w}_t, a_t^h, \bar{w}_t), \qquad w_t = \bar{w}_t + \Delta w_t^{exec}. \tag{5}$$

The feasibility layer respects the HLC direction: increase permits only nonnegative changes, reduce only reductions down to zero, and hold sets the adjustment to zero. It also clips weights to the single-name cap and scales active positive adjustments when buy demand exceeds available cash or the daily turnover budget. The replay buffer stores executed feasible adjustments, and TD3 actor updates use the same layer.

The LLC reward is

$$\begin{aligned} r_t^l = R_t^{net} &- \lambda_{turn}\text{Turnover}_t - \lambda_{dd} d_{t+1} \\ &- \lambda_{risk} \sum_i w_{i,t}\rho_{i,t}, \end{aligned} \tag{6}$$

where R_t^{net} is the post-cost portfolio return. The reward does not assume that low text-risk exposure maximizes raw return; it makes execution trade off return, costs, turnover, drawdown, and exposure to high text-risk assets. Training uses HLC warm-up with directional alignment, LLC warm-up with the HLC frozen, and alternating refinement with increasing execution feedback. All sparse thresholds, risk penalties, and checkpoints are selected on the validation period only. Full implementation details, including reward penalties, training schedules, and feasibility-layer post-processing, are provided in Appendix A.

3. Experiments

Benchmark and protocol. We evaluate HRT on the open FinRL-DeepSeek stock-trading benchmark, constructed from FNSPID financial news and daily market data with precomputed text-derived sentiment and risk signals (Benhenda, 2025; Dong et al., 2024; Liu et al., 2021). The tradable universe is the filtered 89-stock Nasdaq universe released with the benchmark. We use 2013–2018 for training, 2019 for validation, and 2020–2023 once for final testing. This fixed-universe protocol supports reproducibility, but is not a point-in-time dynamic universe; survivorship effects may remain. The 2020–2023 horizon follows the public text-aware benchmark and avoids constructing an additional timestamp-clean news-signal cache.

All strategies use the same 89-stock universe, long-only budget convention, 5% single-name cap, daily weight-based next-open execution, and 10 bps base transaction cost per traded notional. QQQ is reported only as an external Nasdaq-100 proxy. Deterministic baselines are computed once; stochastic learning-based methods report mean $\pm$ seed standard deviation over five seeds. Cumulative return is $\prod_t(1+r_t)-1$, annualized volatility is $\sqrt{252}\sigma_d$, Sharpe is the standard daily-return Sharpe with $r_f = 0$, and turnover is $\sum_i |\Delta w_{i,t}^{exec}|$ without a one-half convention.

Baselines. We compare against same-universe non-RL baselines and learning-based strategies. Equal Weight rebalances uniformly. Minimum Variance uses a rolling covariance estimate with long-only and single-name constraints. Momentum Top-K ranks by past 20-day return. Alpha Top-K ranks by the same predicted forward return $\hat{r}_{i,t}$ used by HRT and holds the top-K names without RL. Flat TD3 is a non-hierarchical continuous-control policy. HRT-Base is a hierarchical PPO–TD3 model using market forecasts and sentiment but without sparse activation, text-risk scores, or the risk-aware LLC reward.

Out-of-sample performance. Table 1 summarizes 2020–2023 performance. HRT achieves the strongest overall return–risk–cost trade-off among learning-based strategies. Compared with HRT-Base, Sharpe rises from 1.06 to 1.24, daily turnover falls from 0.112 to 0.090, and maximum drawdown improves from -0.305 to -0.245. Compared with Alpha Top-K, HRT obtains higher cumulative return with lower drawdown and substantially lower turnover, suggesting that the hierarchy adds value beyond the supervised return forecast. HRT does not dominate every risk metric: Minimum Variance has lower maximum drawdown and CVaR, and HRT-Base has slightly better one-day CVaR. The benefit is therefore a better realized balance among forecast use, risk, and trading frictions. Cumulative-return paths and yearly robustness diagnostics in Appendix B show that this gain is most visible during the 2022 drawdown rather than every bull-market period.

Ablations and execution robustness. Component ablations appear in Appendix B, together with cumulative-return paths, cost-sensitivity curves, trading behavior, and yearly robustness. Sparse activation sacrifices some raw return but improves Sharpe, drawdown, and turnover. Adding text risk lowers ex-post text-risk exposure and improves path-level downside control. The full risk-aware LLC gives the best combined Sharpe, drawdown, turnover, and text-risk exposure. A diagnostic variant without sparse activation reaches higher raw cumulative return, but with worse drawdown and turnover, supporting HRT for risk-adjusted and cost-aware performance rather than maximum raw return.

Table 1. Main out-of-sample performance over 2020–2023. Learning-based methods report mean ± seed standard deviation over five seeds. HRT is not best on every individual risk metric; its advantage is the overall return–risk–cost trade-off among learning-based strategies.

Model	Cum. Ret.	Ann. Ret.	Ann. Vol.	Sharpe	Max DD	CVaR 5%	Turnover
External market proxy							
Market Proxy (QQQ)	0.977	0.186	0.258	0.80	-0.356	-0.031	–
Same-universe non-RL baselines							
Equal Weight	0.643	0.132	0.215	0.70	-0.292	-0.027	0.015
Min-Variance	0.490	0.105	0.155	0.74	**-0.205**	**-0.020**	0.045
Momentum Top-K	0.920	0.177	0.295	0.72	-0.410	-0.037	0.160
Alpha Top-K	1.020	0.192	0.285	0.82	-0.340	-0.033	0.210
Learning-based strategies							
Flat TD3	0.960 ± 0.110	0.183 ± 0.017	0.255 ± 0.018	0.83 ± 0.11	-0.315 ± 0.040	-0.030 ± 0.003	0.195 ± 0.025
HRT-Base	1.151 ± 0.080	0.211 ± 0.011	0.220 ± 0.010	1.06 ± 0.07	-0.305 ± 0.025	-0.027 ± 0.002	0.112 ± 0.010
HRT	$\mathbf{1.321 \pm 0.060}$	$\mathbf{0.234 \pm 0.009}$	0.208 ± 0.010	$\mathbf{1.24 \pm 0.06}$	-0.245 ± 0.015	-0.028 ± 0.003	$\mathbf{0.090 \pm 0.007}$

We also replay policies selected under the 10 bps validation setting under transaction costs from 1 to 50 bps without retraining. All strategies deteriorate as costs rise, but the higher-turnover Alpha Top-K and Flat TD3 baselines degrade more sharply. At 50 bps, HRT retains a positive Sharpe ratio of 0.80, compared with 0.57 for HRT-Base and near-zero values for Alpha Top-K and Flat TD3. Trading-behavior diagnostics in Appendix B further show that HRT trades fewer active names and has lower annualized cost drag than the learning-based alternatives. The hierarchy is therefore useful not only for selecting forecast-consistent trades, but also for avoiding excessive reactions when execution costs rise.

4. Discussion and Conclusions

HRT reframes text-aware trading as forecast-conditioned decision making: structured market forecasts, sentiment scores, and risk signals are evaluated by how well a policy converts them into constrained portfolio decisions. The results support a simple message: separating what to trade from how aggressively to trade can make financial RL more selective and execution-aware.

Several limitations remain. The evaluation follows the filtered 89-stock universe and 2020–2023 horizon supported by the public benchmark, so it tests the proposed hierarchical decision mechanism rather than a production-level live trading system. The backtest uses weight-based fractional execution without integer-share rounding, market impact, liquidity, or capacity constraints. HRT is signal-agnostic, but its performance still depends on the quality, calibration, and timestamp alignment of upstream forecasts and text-derived signals. The factorized HLC is a deliberate scalability trade-off: it avoids the exponential joint action space, but does not fully model all cross-asset dependencies.

Even under these restrictions, the results support the main thesis: forecasting systems should be evaluated not only by predictive quality, but also by how effectively their signals can be converted into constrained decisions. HRT provides one such forecast-to-decision interface. Its hierarchy separates what to trade from how aggressively to trade, turning noisy market and text signals into executable portfolio actions with stronger return–risk–cost behavior.

References

Benhenda, M. FinRL-DeepSeek: LLM-infused risk-sensitive reinforcement learning for trading agents. arXiv preprint arXiv:2502.07393, 2025.

Bertsekas, D. P. *Dynamic Programming and Optimal Control: Volume I*. Athena Scientific, 2012.

Dong, Z., Fan, X., and Peng, Z. FNSPID: A comprehensive financial news dataset in time series. arXiv preprint arXiv:2402.06698, 2024.

Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1587–1596, 2018.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Liu, X.-Y., Yang, H., Gao, J., and Wang, C. D. 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.

Markowitz, H. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Yang, H., Liu, X.-Y., Zhong, S., and Walid, A. Deep reinforcement learning for automated stock trading: An ensemble strategy. In *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020a.

Yang, X., Liu, W., Zhou, D., Bian, J., and Liu, T.-Y. Qlib: An AI-oriented quantitative investment platform. arXiv preprint arXiv:2009.11189, 2020b.

Zhang, B., Yang, H., and Liu, X.-Y. Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. arXiv preprint arXiv:2306.12659, 2023.

Impact Statement

This work studies hierarchical reinforcement learning for decision-aware financial forecasting, where market forecasts and text-derived signals are converted into constrained portfolio actions under transaction-cost, turnover, drawdown, and risk constraints. By separating sparse asset selection from risk-aware execution, HRT offers a structured way to evaluate forecasting systems through downstream decisions rather than point prediction alone. The intended impact is methodological and should not be interpreted as investment advice or evidence of deployable trading performance. Practical deployment would require further validation under liquidity, market impact, compliance, and operational risk constraints.

A. Implementation Details

The feasibility layer in Eq. (5) is implemented as deterministic clipping and scaling. Hold assets are assigned zero adjustment. Reduce assets can only decrease down to zero weight, and increase assets can only use available cash from existing cash and reductions. Desired increases are clipped to the single-name cap and, when buy demand exceeds available cash or the daily turnover budget, positive active adjustments are scaled proportionally. The procedure never applies a final proportional rescaling to all holdings, so hold assets remain unchanged and active adjustments preserve the HLC direction. The replay buffer stores the resulting executed adjustment. Target actions in TD3 are passed through the same feasibility layer, and the layer is treated as deterministic post-processing during actor updates.

Table 2. Implementation configuration. Hyperparameters that affect model selection are chosen using the 2019 validation period only.

Component	**Setting**
Algorithms	PPO for HLC; TD3 for LLC
Market encoder	LightGBM on OHLCV and technical factors
Forecast target	Return from next-open execution to next scheduled rebalance
Risk proxy	20-day rolling realized volatility
Top-K baselines	$K = 30$ for Momentum Top-K and Alpha Top-K
Min-Variance baseline	252-day rolling covariance lookback
Portfolio constraints	Long-only; $w_{\max} = 5\%$; $\tau_{\max} = 0.20$
Sparse threshold	0.58, selected on validation
Transaction cost	10 bps per traded notional
Reward penalties	$\lambda_{turn} = 0.10$, $\lambda_{dd} = 0.05$, $\lambda_{risk} = 0.03$, $\lambda_{act} = 0.01$
HLC reward schedule	$\alpha_t = \alpha_0 \exp(-\lambda t)$, $\alpha_0 = 1.0$, $\lambda = 3 \times 10^{-6}$
Warm-up schedule	HLC: 5×10^4 steps; LLC: 1×10^5 steps
Alternating refinement	One HLC update epoch every five LLC update epochs
Learning rates	PPO: 3×10^{-4}; TD3 actor/critic: 3×10^{-4} / 1×10^{-3}
PPO configuration	Rollout length 2048; 10 epochs; clip ratio 0.20; entropy coefficient 0.01
TD3 configuration	Policy delay 2; target smoothing noise 0.20 clipped at 0.50; soft update 0.005
Replay buffer / batch size	2×10^5 / 256
Discount factor	0.99
Network architecture	Two-layer MLP with 256 hidden units
Training steps	5×10^5
Execution assumption	Weight-based next-open execution; no integer-share rounding
Model selection	Validation-only selection on 2019

B. Additional Results

Table 3. Ablation study. The first four rows add components incrementally; the last removes sparse activation. Values are mean $\pm$ seed standard deviation over five seeds. Text-risk exposure is computed ex post.

Variant	**Cum. Ret.**	**Sharpe**	**Max DD**	**Turnover**	**Text-Risk Exp.**
HRT-Base	1.151 ± 0.080	1.06 ± 0.07	-0.305 ± 0.025	0.112 ± 0.010	0.43 ± 0.02
+ Sparse HLC	1.120 ± 0.075	1.10 ± 0.06	-0.270 ± 0.020	0.101 ± 0.008	0.42 ± 0.02
+ Text Risk Signal	1.205 ± 0.065	1.17 ± 0.06	-0.255 ± 0.018	0.098 ± 0.008	0.39 ± 0.02
+ Risk-Aware LLC (**HRT**)	$\mathbf{1.321 \pm 0.060}$	$\mathbf{1.24 \pm 0.06}$	$\mathbf{-0.245 \pm 0.015}$	$\mathbf{0.090 \pm 0.007}$	$\mathbf{0.35 \pm 0.02}$
HRT w/o Sparse HLC	1.365 ± 0.095	1.12 ± 0.08	-0.335 ± 0.035	0.150 ± 0.018	0.38 ± 0.02

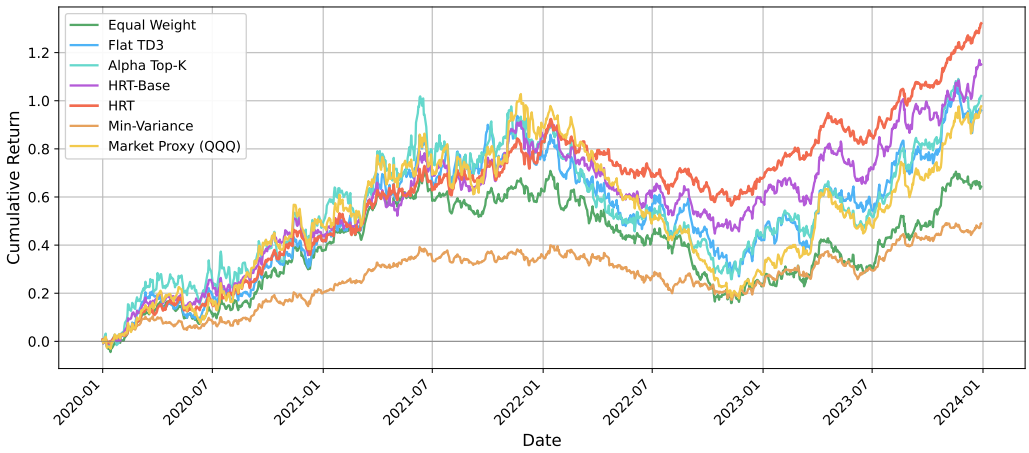

Figure 2. Cumulative returns over 2020–2023. Market Proxy denotes QQQ, an external Nasdaq-100 proxy. HRT reduces downside during the 2022 drawdown and finishes with the strongest return–risk–cost profile among learning-based strategies.

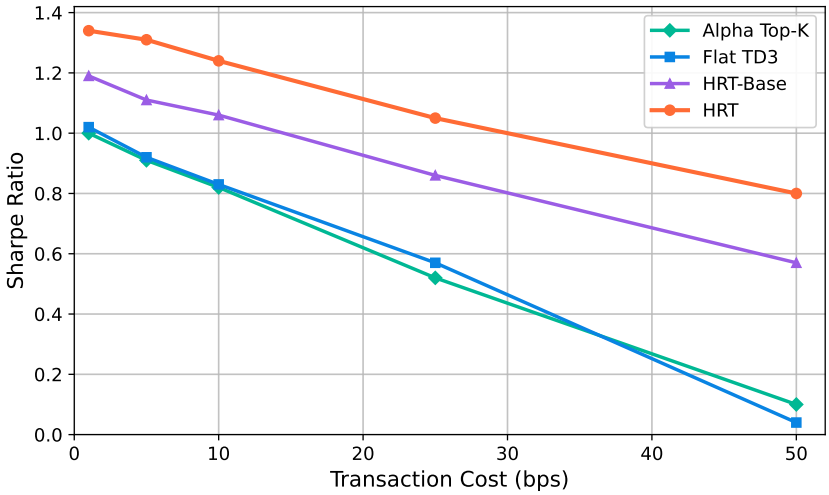

Figure 3. Transaction-cost sensitivity. Policies selected under 10 bps validation costs are replayed under alternative costs without retraining.

Table 4. Trading behavior over 2020–2023. Active ratio is active stocks divided by 89; turnover is $\sum_i |\Delta w_{i,t}^{exec}|$; annualized cost drag is daily turnover $\times$ 10 bps $\times$ 252; HHI is averaged over time; text-risk exposure is computed ex post.

Model	**Active Stocks**	**Active Ratio**	**Turnover**	**Ann. Cost Drag**	**HHI**	**Text-Risk Exp.**
Alpha Top-K	30	0.34	0.210	0.053	0.033	0.48
Flat TD3	40	0.45	0.195	0.049	0.050	0.50
HRT-Base	27	0.30	0.112	0.028	0.044	0.43
HRT	**22**	**0.25**	**0.090**	**0.023**	0.048	**0.35**
HRT w/o Sparse HLC	38	0.43	0.150	0.038	0.041	0.38

Table 5. Yearly robustness analysis. HRT is not designed to dominate the market proxy in every bull-market year; its advantage is most visible in downside control.

Year	**Regime**	**Market Ret.**	**Base Ret.**	**HRT Ret.**	**Base Max DD**	**HRT Max DD**
2020	COVID shock / recovery	0.486	0.440	0.420	-0.285	-0.245
2021	Bull market	0.274	0.290	0.290	-0.110	-0.100
2022	Fed tightening / bear market	-0.326	-0.150	-0.085	-0.305	-0.235
2023	Tech-led recovery	0.549	0.362	0.385	-0.130	-0.115