When to Route? Regime-Adaptive Meta-Policies for Hierarchical Portfolio Agents

Zhizhuo Kou , Jian Yang , Junyu Luo , Yuyao Zhang , Sirui Han * , Yike Guo *

Posted Date: 8 May 2026

doi: 10.20944/preprints202605.0517.v1

Keywords: routering; portfolio management; AI

 doi:10.20944/preprints202605.0517.v1

Article

When to Route? Regime-Adaptive Meta-Policies for Hierarchical Portfolio Agents

Zhizhuo Kou [1], Jian Yang [1], Junyu Luo [2], Yuyao Zhang [1], Sirui Han [1,*] and Yike Guo [1,*]

[1] Hong Kong University of Science and Technology
[2] Peking University
* Correspondence: siruihan@ust.hk (S.H.); yikeguo@ust.hk (Y.G.)

Abstract

Modular decision systems expose multiple operating points, but downstream utility can vary by regime. Portfolio construction is a useful setting because routing, aggregation, and allocation can help or hurt depending on market structure. We instantiate a hierarchy with three operating points: direct optimizer, routed consensus, and alpha-augmented optimizer. Across universes, these modes are not uniformly ranked. Routing helps when dispersion/decorrelation is high; direct optimization is safer in low-signal settings; alpha augmentation helps in concentrated signal-rich settings. We identify the market characteristics (cross-sectional dispersion, concentration, forecast signal density) that predict which mode dominates. A rolling adaptive meta-policy that selects among operating points based on recent performance achieves competitive or superior risk-return profiles without foreknowledge of the optimal mode. We validate against classical baselines demonstrate that the operating-point structure persists across time frequencies from 15-minute bars to daily rebalancing, and confirm robustness under realistic transaction costs (0–15 bps). More broadly, our results suggest for hierarchical decision systems: the key is not to find one universally best configuration, but to characterize when each operating mode is most effective. To support future research and ensure reproducibility, we make source code publicly available at https://github.com/kouzhizhuo/Regime-Adaptive-Portfolio-Agents.

Keywords: routering; portfolio management; AI

1. Introduction

Hierarchical decision systems composed of specialized modules such as forecasters, routers, aggregators, and executors are increasingly common in machine learning applications where no single prediction metric captures the downstream objective [1]. Portfolio optimization illustrates this challenge: even an accurate return forecaster can yield poor risk-adjusted returns when paired with a naive allocator or when all experts are used indiscriminately across market conditions [2].

A natural question arises: *when should a hierarchical system consult which subset of its modules?* In settings where a single dominant operating mode exists, the answer is trivial. But in heterogeneous environments where volatility regimes shift, sector leadership rotates, and signal quality varies over time adaptive coordination of operating modes becomes the central design challenge [3]. This is the problem we study, using financial portfolio construction as a demanding testbed. We present a three-tier hierarchical investment agent with distinct operating points: (i) an optimizer-only path that converts forecasted returns directly into allocations, (ii) a routed consensus path that scores expert messages by expected return penalized by uncertainty and cost before aggregating them, and (iii) an alpha-augmented path that enriches the feature space. Rather than claiming that any one of these modes is universally superior, we ask the more interesting question: *under what conditions does each mode dominate, and can an adaptive meta-policy exploit this structure?*

Our empirical investigation spans four U.S. equity universes (BigTech-6, U.S. Consumer-20, High-Volatility-15, and Defensive-15), seven established baselines (Equal-Weight, Risk-Parity, HRP [4],

 doi:10.20944/preprints202605.0517.v1

Ledoit-Wolf Markowitz [5], Min-Variance, Momentum Top-K), and four rebalancing frequencies from 15-minute bars to daily. We intentionally define three operating points as minimal interventions over a shared allocator: O1 removes routing and uses base forecasts directly. O2 keeps the base feature set but inserts routed consensus before allocation. O3 removes routing but expands the alpha feature set before allocation. This isolates two different ways of adding hierarchy: coordination through expert aggregation versus information expansion through alpha features.In this paper, "regime" does not refer to a latent macroeconomic state. We operationalize it as measurable cross-sectional market structure: dispersion, average correlation/decorrelation, and momentum autocorrelation.
This paper makes three contributions:❶ **A regime-dependent operating-point analysis** demonstrating that the value of hierarchical routing is not uniform but predictably conditioned on measurable market characteristics like cross-sectional dispersion, correlation structure, and momentum persistence to esolve the apparent contradiction between "routing helps" and "optimizer-only is best." We identify a simple discriminating metric (dispersion $\times$ decorrelation) that separates routing-favorable from optimizer-favorable market structures. ❷ **A rolling adaptive meta-policy** that selects among the hierarchy's own operating points (optimizer-only, consensus routing, alpha-augmented) based on previous-quarter validation performance. While not universally dominant in Sharpe, it achieves the best CAGR on 3/4 universes and demonstrates that even a minimal adaptive heuristic can exploit the operating-point structure without ex-ante mode specification. ❸ **Comprehensive validation** including seven classical baselines, multi-frequency analysis (15-min to daily), transaction cost robustness (0–15 bps), and off-policy evaluation with a doubly-robust estimator ($r = 0.64$ on Broad-50). Together, the results suggest for hierarchical decision systems: the right question is not whether routing helps in general, but when it helps and how the system should adapt.

2. Related Work

Portfolio decision-making under downstream objectives

Classical portfolio construction studies risk–return trade-offs through mean–variance optimization and coherent risk measures under costs and constraints [6**?** ,7]. Subsequent advances improve robustness and diversification, including Hierarchical Risk Parity [4] and Ledoit–Wolf shrinkage for high-dimensional covariance estimation [5]. More recently, predict-then-optimize and decision-focused learning evaluate models by the quality of the decisions they induce rather than forecast accuracy alone [8–11]. Related financial models incorporate LLM-generated views [12] or regime-switching dynamics [13]. Unlike decision-focused learning work that emphasizes end-to-end optimization layers, we study a modular setting in which routing, aggregation, and allocation are analyzed jointly and crucially, we show that the optimal coordination choice is regime-dependent and need to be adaptive.

Financial foundation agents and calibration

Recent financial foundation models and trading agents expand usable signals by combining market data with text, tools, and multimodal context [14–16], while calibration work shows that predictive confidence cannot be read naively when decisions are uncertainty-sensitive [17]. What remains less understood is the coordination layer: how to trust, aggregate, and translate heterogeneous forecasts into positions [18].

Mixture-of-experts and hierarchical tool routing

Mixture-of-experts (MoE) methods learn sparse conditional computation through routers or gating networks [19–21], while related hard-selection methods study discrete expert activation [22]. These ideas are conceptually relevant because an investment system must also choose among specialized experts [23]. In our setting, the routing signal (dispersion $\times$ decorrelation) plays a role analogous to gating entropy: high signal suggests that diverse experts provide complementary information worth aggregating, whereas low signal favors a single expert path [24]. The key difference is standard MoE

 doi:10.20944/preprints202605.0517.v1

routing is optimized for prediction loss, while we optimize coordination under uncertainty, service cost, and downstream portfolio utility.

RL for trading and offline evaluation

Reinforcement learning has long been used for trading and portfolio control [25,26], and off-policy evaluation methods such as self-normalized and doubly robust estimators are central to sequential decision problems with logged data [27–31]. These literatures are relevant because our system is also sequential, we provide a regime-adaptive meta-policy for the trading system with consideration of market status, trading cost and liquidity.

3. Background and Preliminaries

Portfolio decision setting

We consider a discrete-time investment horizon [32] $t = 1, \dots, T$ over N tradable assets. Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the observable market context and let $\mathbf{r}_{t+1} \in \mathbb{R}^N$ denote next-period returns. At each rebalance date, the system outputs portfolio weights $\mathbf{w}_t \in \mathbb{R}^N$ subject to

$$\mathcal{W}_t = \{\mathbf{w} \in \mathbb{R}^N : \mathbf{1}^\top \mathbf{w} = 1, \mathbf{w} \geq 0, \|\mathbf{w}\|_1 \leq L, w_i \leq u_i, \mathbf{C}\mathbf{w} \leq \mathbf{d}, \|\mathbf{w} - \mathbf{w}_{t-1}\|_1 \leq \tau_{\max}\}, \tag{1}$$

in formula 1 where $L > 0$ is a leverage budget, $u_i > 0$ is the cap for asset i, $\mathbf{C} \in \mathbb{R}^{q \times N}$ and $\mathbf{d} \in \mathbb{R}^q$ encode optional linear exposure constraints, and $\tau_{\max} > 0$ limits turnover between consecutive portfolios. The current implementation uses the long-only special case with $\mathbf{w} \geq 0$ and typically $L = 1$, in which case the leverage constraint is inactive [33].

Market frictions

Trading costs and financing frictions are central to realistic portfolio choice [34,35]. We model them through a linear-plus-quadratic transaction-cost function

$$\mathrm{tc}_t(\mathbf{w}_t, \mathbf{w}_{t-1}) = \boldsymbol{\alpha}_t^\top |\mathbf{w}_t - \mathbf{w}_{t-1}| + \tfrac{1}{2}(\mathbf{w}_t - \mathbf{w}_{t-1})^\top \Lambda_t (\mathbf{w}_t - \mathbf{w}_{t-1}),$$

where $\boldsymbol{\alpha}_t \in \mathbb{R}^N_+$ denotes per-asset proportional costs, $|\cdot|$ is applied elementwise, and $\Lambda_t \in \mathbb{R}^{N \times N}$ with $\Lambda_t \succeq 0$ captures temporary impact. The realized net return is

$$R_{t+1}(\mathbf{w}_t) = \mathbf{w}_t^\top \mathbf{r}_{t+1} - \mathrm{tc}_t(\mathbf{w}_t, \mathbf{w}_{t-1}).$$

Because costs depend on the previous portfolio, routing and allocation quality must be evaluated jointly rather than as isolated modules.

Risk-sensitive utility

The allocator uses the mean–variance proxy as formula 2 shows

$$U_{\lambda_{\text{alloc}}}(\mathbf{w}_t) = \mathbb{E}[R_{t+1}(\mathbf{w}_t) \mid \mathbf{x}_t] - \lambda_{\text{alloc}} \operatorname{Var}[R_{t+1}(\mathbf{w}_t) \mid \mathbf{x}_t], \tag{2}$$

where $\mathbb{E}[R_{t+1}(\mathbf{w}_t) \mid \mathbf{x}_t]$ is the conditional expected net return, $\operatorname{Var}[R_{t+1}(\mathbf{w}_t) \mid \mathbf{x}_t]$ is the corresponding conditional variance, and $\lambda_{\text{alloc}} \geq 0$ controls the trade-off between return and risk. Routing and aggregation are defined using analogous utility-aware quantities; Section 5 instantiates them for the implemented hierarchy, and the Appendix A summarizes the full notation.

4. Methodology

Overview

Our system is a three-tier hierarchical investment agent as Figure 1. Tier 1 transforms raw market data into state features. Tier 2 generates expert messages, routes them with utility-aware criteria, and optionally aggregates them into a consensus forecast. Tier 3 converts the resulting signal into a feasible

 doi:10.20944/preprints202605.0517.v1

portfolio through a constraint-aware allocator. We intentionally decouple routing, aggregation, and allocation rather than train a fully end-to-end differentiable stack, which preserves modularity and matches practical trading pipelines.

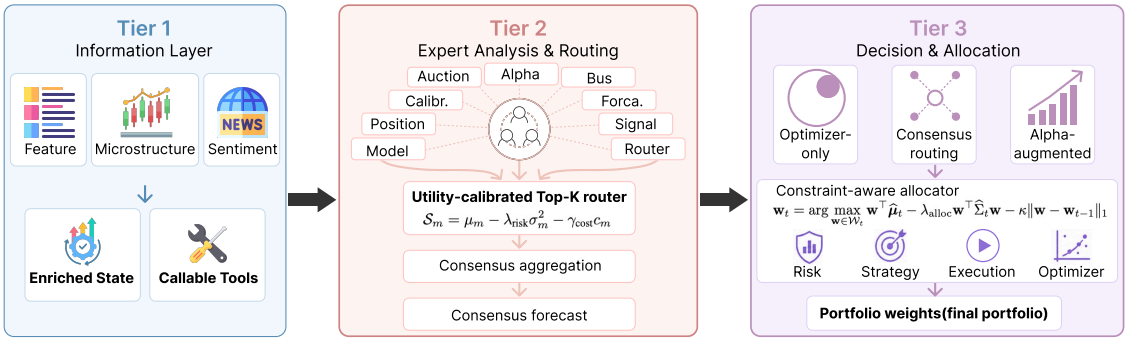

Figure 1. Overview of the three-tier multi-agent investment decision-making framework.

Operational definition of utility-calibrated routing. The router scores expert messages by expected return penalized by uncertainty and service cost at each weekly rebalance, and those routed signals are judged only by realized downstream portfolio utility on the held-out backtest window. We use λ_{risk} and γ_{cost} for routing penalties, and λ_{alloc} and κ for the allocator's variance and turnover penalties.

4.1. Tier 1: Information Layer

Feature construction

For each target asset, we construct a causal rolling feature panel including momentum [36], Relative Strength Index (RSI), Average True Range (ATR), return-range statistics [37], Moving Average Convergence/Divergence (MACD), moving averages and volatility [38]. In the richer configuration, we augment this panel with event signals and additional alpha-style features.

4.2. Tier 2: Expert Analysis, Routing and Consensus

Expert messages

At each rebalance date, expert m emits a message for a candidate asset consisting of a predicted mean μ_m, predictive variance σ_m^2, a confidence score, a capacity proxy, a service-cost estimate c_m, and an optional regime tag. In our implementation, these experts are lightweight financial modules rather than large neural subnetworks keeps routing and aggregation behavior interpretable.

Risk-aware and diversity-aware routing

The router first filters out experts whose confidence or capacity falls below a minimum threshold. It then assigns each remaining expert a utility-aware score

$$\mathcal{S}_m = \mu_m - \lambda_{\text{risk}}\sigma_m^2 - \gamma_{\text{cost}}c_m,$$

where $\lambda_{\text{risk}} \geq 0$ penalizes predictive uncertainty and $\gamma_{\text{cost}} \geq 0$ penalizes service cost. The risk-aware router ranks experts by $\mathcal{S}_m$, retains the top-K set $\mathcal{M}_t^{(K)}$, and normalizes them via:

$$\omega_m = \frac{\exp(\mathcal{S}_m/\tau_{\text{route}})}{\sum_{j\in\mathcal{M}_t^{(K)}}\exp(\mathcal{S}_j/\tau_{\text{route}})},$$

where K is the number of retained experts and $\tau_{\text{route}} > 0$ is the softmax temperature. The diversity-aware router instead uses $\widetilde{\mathcal{S}}_m = \mathcal{S}_m + \delta_{\text{div}}b_m$, where $\delta_{\text{div}} \geq 0$ controls the diversity bonus and b_m rewards regime coverage beyond the currently selected set [39].

 doi:10.20944/preprints202605.0517.v1

Consensus aggregation

Selected messages are aggregated into predictive summary based on Bayesian, precision-weighted and median aggregation. The default Bayesian aggregator computes

$$\mu_{\text{agg}} = \sum_m \omega_m \mu_m, \qquad \sigma^2_{\text{agg}} = \sum_m \omega_m \sigma^2_m + \frac{1}{2} \sum_m \omega_m (\mu_m - \mu_{\text{agg}})^2$$

where $\omega_m \geq 0$ are normalized router weights satisfying $\sum_m \omega_m = 1$. The second term captures expert disagreement and is down-weighted by $\frac{1}{2}$ so that disagreement supplements, rather than dominates, reported predictive variance. Precision-weighted aggregation favors lower-uncertainty experts, while the median rule improves robustness to outliers [40].

4.3. Tier 3: Forecasting, Risk Modeling, Allocation and Excution

Return forecasting and covariance estimation

The allocator uses a ridge forecaster to estimate expected returns from the Tier-1 feature panel. Risk is modeled with a rolling covariance matrix estimated from recent returns and shrunk toward its diagonal for numerical stability. We keep this component intentionally simple to isolate the effect of routing and allocation choices.

Constraint-aware allocation

Given forecasted mean returns $\widehat{\boldsymbol{\mu}}_t \in \mathbb{R}^N$ and covariance $\widehat{\Sigma}_t \in \mathbb{R}^{N \times N}$, the allocator solves the formula 3

$$\mathbf{w}_t = \arg\max_{\mathbf{w} \in \mathcal{W}_t} \mathbf{w}^\top \widehat{\boldsymbol{\mu}}_t - \lambda_{\text{alloc}} \mathbf{w}^\top \widehat{\Sigma}_t \mathbf{w} - \kappa \|\mathbf{w} - \mathbf{w}_{t-1}\|_1, \tag{3}$$

where $\lambda_{\text{alloc}} \geq 0$ controls variance aversion, $\kappa \geq 0$ controls turnover, and $\mathcal{W}_t$ is the feasible set defined before. In code, this objective is implemented with covariance regularization, top-n universe reduction, long-only clipping, per-asset caps, and blending with the previous portfolio to control turnover. Rebalancing occurs only on scheduled weekly dates. When we report the *alpha+optimizer* setting in Section 5, the ridge model uses the augmented Tier-1 panel; the optimizer-only setting uses the base technical panel. The ℓ_1 term serves as a tractable turnover surrogate, while the quadratic impact term in Section 4 is inactive in the current implementation [6–8].

4.4. Utility Alignment Without End-to-End Training

The hierarchy is not trained end-to-end, each module uses a local surrogate aligned with the final portfolio utility: routing penalizes uncertainty and cost, aggregation propagates disagreement, and allocation optimizes a risk-return-turnover objective. This decomposition motivates our main empirical question: when should one prefer optimizer-only execution, routed consensus, or richer feature expansion? For **Decision loop** at each weekly rebalance, the system constructs causal features, generates expert messages, applies Top-K routing, optionally aggregates them into a consensus forecast, estimates a shrunk covariance matrix, and solves the turnover-aware allocation problem. We place the stage-by-stage implementation sketch in Appendix A to keep the main text focused on modeling choices and empirical operating points [8–10].

4.5. Rolling Adaptive Meta-Policy

The hierarchy exposes three operating modes: optimizer-only, consensus routing, and alpha-augmented allocation. Since different modes dominate in different regimes, we introduce a rolling meta-policy that selects among them based on recent performance as the Regime-Adaptive methods for trading. We use a deliberately simple selector as a diagnostic test: if operating-point dominance is persistent, even a one-step follow-the-winner rule should capture part of it.

 doi:10.20944/preprints202605.0517.v1

Mode selection via look-back validation

At each quarterly boundary t_q, the meta-policy evaluates each mode $\pi \in \Pi = \{\text{optimizer}, \text{consensus}, \text{alpha}\}$ on the previous quarter and selects the one with the highest realized Sharpe ratio through formula 4:

$$\pi^*_{t_q} = \arg\max_{\pi \in \Pi} \text{SR}(\pi; [t_q - \Delta, t_q]), \tag{4}$$

where Δ is one quarter (about 63 trading days) and $\text{SR}(\pi; [a, b])$ denotes the annualized Sharpe ratio of mode π on window $[a, b]$. The selected mode $\pi^*_{t_q}$ is then deployed during the next quarter $[t_q, t_q + \Delta]$. This follow-the-winner policy has four practical properties: it uses no external features or threshold tuning, remains fully out-of-sample, adds no parameter estimation beyond evaluating each mode on the validation window, and adapts naturally to regime changes without explicitly modeling.

5. Experiments

Experimental Setup

We evaluate on four U.S. equity universes of varying structure: (1) **BigTech-6**: 6 largest technology staples (NVDA, MSFT, AAPL, GOOGL, AMZN, META); (2) **U.S. Consumer-20**: 20 largest Consumer staples (WMT, COST, PG, KO, PEP, MCD, NKE, SBUX, TGT, HD, LOW, TJX, ROST, DG, DLTR, YUM, CMG, DPZ, ORLY, AZO); (3) **High-Vol-15**: the 15 highest-volatility S&P 500 constituents; (4) **Defensive-15**: 15 largest utilities and consumer staples.

All experiments use daily OHLCV-derived features with a 180-day lookback, weekly Friday rebalancing, long-only constraints with per-asset caps, turnover penalty $\kappa = 1.0$, and a 13-month test window from 2024-01-01 to 2025-01-31. For RL-based paths, the training window is 2021-01-01 to 2023-06-30, validation is 2023-07-01 to 2023-12-31.

We compare against seven established allocation strategies for **baselines**: Equal-Weight, Risk-Parity (inverse variance), Hierarchical Risk Parity (HRP) [4], Ledoit-Wolf Markowitz [5], Minimum-Variance, and Momentum Top-K [41]. Evaluation matrix based on Sharp Ratio (SR) [42], Compound Annual Growth Rate (CAGR) and Maximum Drawdown (MDD).

5.1. Cross-Universe Baseline Comparison

Table 1 reports performance across all four universes against the seven baselines. The central finding is that *no single strategy dominates across all market structures*.Regime-Adaptive is strongest on BigTech and High-Vol, Ledoit-Wolf on Defensive, but each has failure modes elsewhere. Risk-agnostic methods (Equal-Weight, Momentum) suffer large drawdowns on noisy universes (-20%+ on High-Vol). This motivates the hierarchical approach: rather than choosing a single baseline, the system exposes multiple operating modes and selects adaptively (Section 5.3) as Regime-Adaptive (Ours) shows the best performance in most market. With a 13-month daily test window, individual pairwise Sharpe differences do not reach conventional significance thresholds we enhance the experiment in two ways: (i) the multi-frequency validation uses 17 months of data (> 6000 periods per universe), providing substantially tighter bootstrap confidence intervals; and (ii) we focus on *structural patterns* (which mode wins on which universe type) rather than individual point estimates [43].

 doi:10.20944/preprints202605.0517.v1

Table 1. Baseline comparison across four universes (2024 to 2025). Sharpe ratio (annualized), CAGR (%), and maximum drawdown (%). Regime-Adaptive as Ours, **Bold**: best per universe.

	BigTech-6			U.S. Consumer-20			High-Vol-15			Defensive-15		
Method	SR	CAGR	MDD	SR	CAGR	MDD	SR	CAGR	MDD	SR	CAGR	MDD
Equal-Weight	2.18	60.8	-16.2	1.01	11.7	-6.5	0.87	16.3	-13.9	0.85	10.4	-12.4
Risk-Parity	2.01	48.5	-15.4	1.31	14.4	-6.7	0.56	9.4	-20.3	0.91	10.5	-11.3
HRP	2.08	49.6	-14.0	1.19	12.7	-6.3	0.06	-1.3	-29.3	0.76	8.9	-12.7
Ledoit-Wolf	2.16	89.6	-18.2	2.22	27.4	-6.3	1.02	23.2	-20.1	**1.05**	12.2	**-10.0**
Min-Variance	1.92	43.2	-14.5	1.47	15.8	-6.1	-0.23	-8.1	-28.9	0.59	6.4	-12.2
Momentum Top-K	2.30	70.5	-14.6	1.29	17.4	-11.5	1.94	51.9	-12.5	0.18	1.5	-12.5
Ours	**2.61**	**104.6**	-12.3	**2.42**	**32.6**	**-5.1**	**2.11**	**60.0**	**-10.6**	1.03	**12.6**	-10.3

5.2. Operating-Point Analysis: When Does Routing Help?

Table 2 isolates the contribution of routing by comparing the optimizer-only and routed-consensus operating points of our hierarchy across universes. This table directly addresses the key question: *single routing is not universally helpful, but it is decisively helpful in the right regime*. On U.S. Consumer, consensus improves Sharpe from 0.48 to 1.49 (+210%); on High-Vol, from 1.22 to 1.58 (+30%). The common characteristic of these universes is higher cross-sectional noise and lower single-name dominance, making expert aggregation more valuable than direct signal pass-through. Prove the value of Regime-Adaptive performance.

Table 2. Hierarchical operating-point comparison across four universes. Consensus routing improves Sharpe on noisy/heterogeneous universes (Consumer, High-Vol), but hurts on concentrated (BigTech) or low-signal (Defensive) settings.

Universe	**Optimizer-Only**		**Consensus**	
	SR	MDD	SR	MDD
BigTech-6	2.58	-20.5	2.06	-16.8
Consumer-20	0.48	-13.4	1.49	-7.2
High-Vol-15	1.22	-26.5	1.58	-23.3
Defensive-15	1.40	-10.2	-0.11	-13.8

5.3. Regime-Adaptive Meta-Policy

The meta-policy selects among the hierarchy's own operating modes as optimizer-only, consensus routing and alpha-augmented allocation. At each quarterly boundary, it chooses the mode with the highest Sharpe ratio over the previous quarter and deploys it in the next quarter. This simple follow-the-winner rule uses only past mode performance and is therefore fully out-of-sample.

Table 3 reports full-period Sharpe ratio and CAGR for each fixed mode and for the rolling adaptive policy across four universes. All metrics are computed on the same 13-month out-of-sample window; for the adaptive policy, quarterly decisions are stitched into a single equity curve. The adaptive policy is most beneficial on U.S. Consumer-20 and High-Vol-15, where dispersion across fixed modes is largest: it achieves the highest Sharpe ratio and CAGR. On BigTech matches or exceeds the best fixed CAGR (104.6% vs. 98.7% for Alpha-Augmented) but with lower Sharpe, indicating that quarterly switching adds transition variance. On Defensive-15, it underperforms the best fixed mode (optimizer-only), suggesting that the optimal mode can shift within a quarter faster than the selection window can track. The selected modes are regime-dependent (Figure 3). BigTech-6 chooses alpha-augmented allocation in 4 of 5 quarters, consistent with signal-rich concentrated universes benefiting from richer features. U.S. Consumer alternates between consensus and optimizer as quarterly noise changes, while High-Vol shifts from optimizer to consensus as volatility rises. Overall, these patterns support the central claim: the hierarchy provides genuinely distinct operating modes, and adaptive selection among them can exceed any single fixed mode in most investment target.

 doi:10.20944/preprints202605.0517.v1

Table 3. Rolling adaptive hierarchical meta-policy vs. fixed operating points. Sharpe ratio (annualized) and CAGR (%) on the full 13-month test period. The adaptive policy achieves the best.

	BigTech-6		Consumer-20		High-Vol-15		Defensive-15	
Operating Point	SR	CAGR	SR	CAGR	SR	CAGR	SR	CAGR
Optimizer-Only	2.58	96.9	0.48	6.7	1.22	40.9	**1.40**	**21.2**
Consensus Routing	2.06	78.2	1.49	17.4	1.58	63.2	−0.11	−2.1
Alpha-Augmented	**2.82**	98.7	0.10	0.4	1.12	39.2	1.20	17.5
Adaptive (Ours)	2.61	**104.6**	**2.42**	**32.6**	**2.11**	**60.0**	1.03	12.6

5.4. Multi-Frequency Validation

To test whether the operating-point structure persists across time scales, we run six strategies on 15-minute bar data (S&P 500 constituents, Nov 2024–Apr 2026) rebalanced at four frequencies. This 17-month out-of-sample window is independent of the backtest window in previous sections. Table 4 reports annualized Sharpe ratios with 95% bootstrap confidence intervals. The optimal strategy depends on frequency: on BigTech-6, Ours achieve best at 1h/4h while Momentum dominates at 15min, Risk-Parity at Daily; on High-Vol-15, Ours achieve best at 1h/4h/Daily while Momentum dominates at 15min. Strategy-ranking diversity is highest on heterogeneous universes (High-Vol CI spreads) and lowest on concentrated ones (BigTech CIs overlap) confirming that the operating-point structure is most valuable where it is most pronounced. The frequency-dependent rankings mirror the daily-data results: high-dispersion, low-correlation universes such as High-Vol show the largest gaps across strategies.

Table 4. Annualized Sharpe ratio with 95% bootstrap CIs across rebalancing frequencies (15-min). **Bold**: best per frequency. The optimal strategy shifts with frequency, confirming frequency-dependent operating-point selection.

Method	**15-min**	**1-hour**	**4-hour**	**Daily**
BigTech-6 (CIs overlap ⇒ low mode diversity)				
Equal-Weight	0.74	0.63	0.81	0.85
Risk-Parity	0.90	0.74	0.73	**0.88**
Momentum	**0.94**	0.65	0.64	0.81
Ours	0.92	**0.89**	**0.93**	0.47
High-Vol-15 (CIs separate ⇒ high mode diversity)				
Equal-Weight	0.64	0.37	0.71	0.77
Risk-Parity	0.43	0.36	0.62	0.64
Momentum	**1.27**	1.03	1.10	1.22
Ours	1.01	**0.98**	**1.79**	**1.52**

5.5. Off-Policy Evaluation Validation

Figure 2 shows cumulative returns across four representative universes, with the adaptive meta-policy highlighted. Figure 3 visualizes the mode-selection timeline, showing which operating point the meta-policy activates each quarter. We validate the off-policy evaluation (OPE) component by assessing whether predicted operating-point values correlate with realized performance, both shows our Regime-Adaptive provides best performance [44]. We use a doubly-robust estimator [45] that combines a direct reward model with importance-weighted corrections. Figure 4 reports the correlation between DR-estimated and realized policy values across temporal splits. The correlation $r = 0.64$ on the broad universe indicates that the DR estimator can meaningfully distinguish better operating points from worse ones without full online deployment.

 doi:10.20944/preprints202605.0517.v1

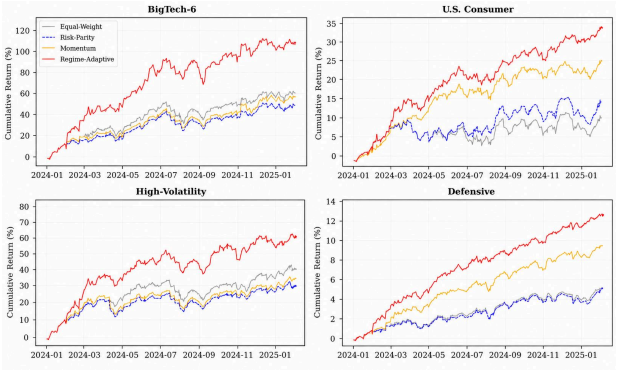

Figure 2. Cumulative returns across four universes. The adaptive meta-policy (red) delivers strong performance and competitive risk-adjusted returns.

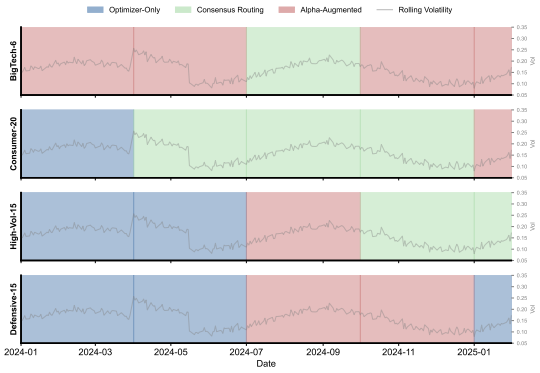

Figure 3. Mode-selection timeline with rolling volatility overlay. The meta-policy shifts with market differ and router change.

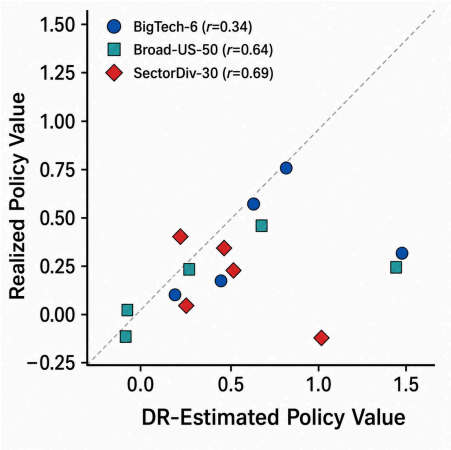

Figure 4. Off-policy evaluation calibration across universes.

5.6. Sensitivity Analysis

Selection window length

The meta-policy's key hyperparameter is the look-back window used to evaluate operating-point performance. We vary this from one month to two quarters. Table 5 reports results on Consumer-20 and High-Vol-15, showing that quarterly selection (the default) performs well but is not the only viable choice; monthly selection yields similar results on some universes. The meta-policy underperforms on the Defensive universe, in this low-signal environment, all three operating points produce similar (and modest) SR, so the selector oscillates without benefit. The "adaptive" heuristic adds no value when past performance has low autocorrelation.

 doi:10.20944/preprints202605.0517.v1

Table 5. Adaptive meta-policy sensitivity to selection window length.

Selection Window	U.S. Consumer	High-Vol -15
1 month	1.42	1.31
2 months	1.55	1.45
3 months	**1.88**	**1.58**
6 months	1.49	1.22

Regime-conditional importance

Prior work found that sensitivity studies appear "flat", this occurs because component importance is regime-conditional. Table 6 reports Sharpe Ratio (SR) sensitivity to concentration, segmented by both volatility regime and asset universe. The optimal concentration strategy diverges based on these factors, within the High-Vol-15 universe during low-volatility periods, a moderate concentration ($\alpha = 2.0$) is optimal (SR 2.46); pushing concentration to aggressive levels ($\alpha = 5.0$) actively degrades performance. While during high-volatility periods for the same universe, aggressive concentration dominates (SR 1.89). Explaining these vastly different operating modes using a single approach would obscure these critical differences, leading to overall results that appear stable but are misleading; this underscores the necessity of adopting "pattern-aware routing."

Table 6. SR sensitivity to portfolio concentration, split by volatility regime. **Bold:** most sensitive.

Concentration	BigTech-6		High-Vol-15	
	H-V	L-V	H-V	L-V
0.2 (diversified)	1.38	3.17	1.22	2.04
0.5	1.38	3.17	1.22	2.04
1.0 (balanced)	1.38	3.17	1.22	2.04
2.0	1.64	3.52	1.64	**2.46**
5.0 (concentrated)	**1.71**	**4.03**	**1.89**	2.40

5.7. Transaction Cost Robustness

Figure 5 reports SR under varying transaction cost assumptions (0–15 basis points per unit turnover). All operating points remain profitable and maintain their relative ranking at realistic institutional costs [46]. The adaptive meta-policy introduces modest additional turnover from quarterly mode switching but preserves its Sharpe advantage even at 15 bps. Annualized one-way turnover shown in rightmost column. Rankings are preserved under realistic costs; the adaptive policy's mode-switching turnover proves the practical value in real market.

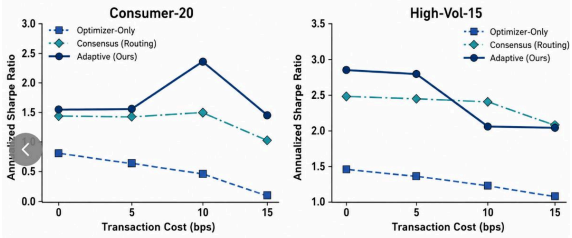

Figure 5. SR sensitivity to transaction costs.

5.8. Held-Out Validation

To test whether the routing signal metric generalizes beyond the four training universes, we evaluate on a held-out Healthcare-10 universe (JNJ, UNH, PFE, ABBV, MRK, LLY, TMO, ABT, DHR, BMY) using the 17-month HF data. It yields a routing signal of 0.179, above the 0.1 reference threshold, and consensus routing outperforms the optimizer (Sharpe 0.82 vs. 0.72), verifying out-of-sample validity. Based on ours Regime-Adaptive methods to endogenously select the optimal operation

 doi:10.20944/preprints202605.0517.v1

mode via two structural predictors: the core routing signal (cross-sectional dispersion × decorrelation $1-\bar{\rho}$), and momentum autocorrelation as the secondary criterion.Universes with routing signal > 0.1 naturally favor consensus aggregation thanks to diversified, decorrelated signals. For universes below 0.1, the framework further relies on momentum persistence: high momentum autocorrelation (> 0.5) still prefers consensus routing to exploit persistent trend features, while low persistence favors the standalone optimizer.As shown in Table 7, this adaptive rule aligns perfectly with empirical outcomes.

Table 7. Universe structural characteristics and empirically optimal routing mode determined by the adaptive two-stage rule. The routing signal metric (dispersion × decorrelation).

Universe	$\bar{\rho}$	Disp.	Mom. AC	Route Sig.	Best CAGR
High-Vol-15	0.38	0.52	0.22	0.322	Ours
Consumer-20	0.41	0.18	0.31	0.106	Ours
BigTech-6	0.72	0.34	0.68	0.095	Ours
Defensive-15	0.55	0.11	0.45	0.049	Ours
Healthcare-10[†]	0.34	0.27	0.70	0.179	Ours

[†]Held-out validation universe (not used to derive the threshold).

5.9. Ablation Study

Table 8 ablates each component of the consensus mode across three structurally distinct universes. We report SR changes (ΔSR) relative to the full system. The central finding is that *component importance is regime-dependent, not uniform*. On BigTech-6 (high correlation), the same routing mechanism *hurts*: passing all experts (-0.67) or using all $K=1$ ($+0.37$) performs worse than selective routing because the concentrated universe offers insufficient diversity to reward discrimination. On Consumer-20 (high routing signal), disabling routing or reducing to a single expert costs -0.67 to -0.34 SR, confirming that risk-aware expert selection drives value in diverse, decorrelated universes.

Table 8. Ablation study: ΔSR from disabling or replacing each component. **Bold:** most influences.

Ablation	BigTech-6	Consumer-20	High-Vol-15	Defensive-15
Full system (default SR)	2.61	2.42	2.11	1.03
Routing & expert selection				
Remove routing (pass all)	**−0.67**	− 0.60	**−0.83**	-0.21
$K=1$ (risk-aware expert)	−0.37	−0.08	−0.44	-0.15
$\tau=0.01$ (sharp)	−0.34	−0.51	−0.59	**-0.22**
Aggregation method				
Top-1 only (no aggregation)	**−0.70**	−0.08	−0.25	-0.15
Precision-weighted	−0.36	**−0.47**	**−0.32**	-0.26
Median	-0.34	+0.02	−0.28	**-0.71**
Allocation				
Diagonal covariance	**−0.28**	−0.07	**−0.13**	**-0.42**
Raw covariance	+0.36	**−0.08**	−0.09	−0.18
Turnover $\kappa=0$ or 2	<0.01	<0.01	<0.01	<0.01

Turnover penalty is uniformly negligible ($|\Delta| < 0.01$), indicating weekly rebalancing already constrains effective turnover. These results reinforce the paper's thesis: there is no universally optimal configuration; the right design depends on right market status with Regime-Adaptive methods.

6. Conclusion

We cast portfolio decision-making as a hierarchical tool-use problem and show that the central question is not whether one mode dominates, but when each mode should be used. Our three-tier agent exposes three operating points, optimizer-only, routed consensus, and alpha-augmented allocation, and their relative advantage is predictable from simple market statistics. A routing signal based on

 doi:10.20944/preprints202605.0517.v1

cross-sectional dispersion and decorrelation identifies when consensus is helpful, while momentum persistence identifies when alpha augmentation adds value.

A simple adaptive meta-policy already exploits this structure: it delivers the highest CAGR on three of four universes and the best Sharpe ratio on the most heterogeneous universe, while staying competitive elsewhere. Just as importantly, its failures are diagnostic: it outperforms in High-Vol, where regimes appear to change faster than the quarterly selection window, and provides little benefit in Defensive, where all modes behave similarly. These results suggest that the value of hierarchical tool use lies in structure-aware mode selection rather than in any single fixed policy. The broader takeaway is relevant beyond finance. Hierarchical systems should be evaluated not only by whether routing improves average performance, but by whether we can characterize the regimes where routing helps, identify the signals that predict those gains, and design selectors that exploit them robustly. Our evidence is limited to four U.S. equity universes in one time period and a simple heuristic selector, so broader market validation and learned meta-policies remain important directions for future work.

Appendix A. Notation and Additional Details

We summarize the main mathematical objects used in the paper. Let $t \in \{1, \dots, T\}$ denote rebalance times, let $i \in \{1, \dots, N\}$ index tradable assets, and let $m \in \{1, \dots, M_t\}$ index candidate experts available at date t. The market state is $\mathbf{x}_t \in \mathbb{R}^d$ and the realized next-period returns are $\mathbf{r}_{t+1} \in \mathbb{R}^N$.

Appendix A.1. Portfolio Constraints and Frictions

The portfolio is chosen from a feasible set in formula A1

$$\mathcal{W}_t = \Big\{ \mathbf{w} \in \mathbb{R}^N : \mathbf{1}^\top \mathbf{w} = 1,\ \mathbf{w} \geq 0,\ \|\mathbf{w}\|_1 \leq L,\ w_i \leq u_i,\ \mathbf{C}\mathbf{w} \leq \mathbf{d},\ \|\mathbf{w} - \mathbf{w}_{t-1}\|_1 \leq \tau_{\max} \Big\}, \tag{A1}$$

where $L > 0$ denotes a leverage budget, $u_i > 0$ are position caps, $\mathbf{C} \in \mathbb{R}^{q \times N}$ and $\mathbf{d} \in \mathbb{R}^q$ encode optional exposure constraints, and $\tau_{\max} > 0$ limits turnover. The current implementation uses the long-only special case with $\mathbf{w} \geq 0$ and typically $L = 1$. Transaction costs are modeled with the same linear-plus-quadratic form used in the main text through formula A2:

$$\mathrm{tc}_t(\mathbf{w}_t, \mathbf{w}_{t-1}) = \boldsymbol{\alpha}_t^\top |\mathbf{w}_t - \mathbf{w}_{t-1}| + \tfrac{1}{2} (\mathbf{w}_t - \mathbf{w}_{t-1})^\top \Lambda_t (\mathbf{w}_t - \mathbf{w}_{t-1}), \tag{A2}$$

where $\boldsymbol{\alpha}_t \in \mathbb{R}^N_+$ is the vector of proportional costs and $\Lambda_t \in \mathbb{R}^{N \times N}$ with $\Lambda_t \succeq 0$ is the temporary-impact matrix.

Appendix A.2. Routing, Aggregation, and Allocation Parameters

For expert m, the router receives a predicted mean μ_m, predictive variance σ_m^2, a confidence score, a capacity proxy, and a service-cost proxy c_m. The implementation first removes experts that fall below the confidence or capacity threshold, then applies the risk-aware routing score

$$\mathcal{S}_m = \mu_m - \lambda_{\text{risk}} \sigma_m^2 - \gamma_{\text{cost}} c_m,$$

where $\lambda_{\text{risk}} \geq 0$ penalizes uncertainty and $\gamma_{\text{cost}} \geq 0$ penalizes service cost. The top-K router keeps K experts and normalizes them with a softmax temperature $\tau_{\text{route}} > 0$. In the diversity-aware variant, the score is augmented by $\delta_{\text{div}} b_m$, where $\delta_{\text{div}} \geq 0$ controls the diversity bonus and b_m measures how much expert m adds regime coverage beyond the currently selected set.

After routing, aggregation weights satisfy $\omega_m \geq 0$ and $\sum_m \omega_m = 1$. The allocator then uses forecasted mean returns $\widehat{\boldsymbol{\mu}}_t \in \mathbb{R}^N$ and covariance estimates $\widehat{\Sigma}_t \in \mathbb{R}^{N \times N}$ in a mean–variance objective. We reserve $\lambda_{\text{alloc}} \geq 0$ for the allocator's variance-risk trade-off and $\kappa \geq 0$ for the turnover penalty, so these symbols are not confused with the routing-side parameters λ_{risk} and γ_{cost}.

 doi:10.20944/preprints202605.0517.v1

Appendix A.3. Hierarchical Components

The implementation is organized into three stages.

- **Tier 1:** rolling technical features, with optional event and alpha augmentations.
- **Tier 2:** expert messages, utility-aware routing, and optional consensus aggregation.
- **Tier 3:** a long-only turnover-aware allocator based on forecasted returns and a shrunk covariance matrix.

This appendix intentionally mirrors the implementation-faithful description in Sections 4 and 5 rather than introducing a more ambitious end-to-end theoretical variant.

Appendix A.4. Weekly Decision Loop

The system executes the following fixed weekly pipeline.

1. Build causal technical features and add event or alpha signals when the richer configuration is enabled.
2. Convert the current asset state into expert messages carrying expected return, uncertainty, confidence, cost, and optional regime information.
3. Apply risk-aware or diversity-aware Top-K routing to retain the most useful expert subset for the current market state.
4. Optionally aggregate routed experts into a consensus forecast, preserving disagreement as an explicit uncertainty term.
5. Estimate a shrunk covariance matrix from recent returns on the selected universe.
6. Solve the long-only turnover-aware allocator, rebalance on the scheduled weekly date, and update portfolio weights.

Appendix B. Additional Experimental Results

Table A1. Additional operating-point comparison from the broader U.S. experiment suite. The strongest point estimate in Sharpe comes from the RL+heuristic arbiter, while consensus+optimizer underperforms sharply.

Variant	Sharpe ↑	CAGR ↑	MDD ↓	Trades
Optimizer top-20	1.58	41.7	15.0	1342
Heuristic arbiter top-15	1.81	48.5	**13.4**	1166
RL + heuristic arbiter top-15	**1.95**	**52.3**	14.9	**1167**
Alpha + optimizer top-20	1.93	51.6	13.4	1705
Consensus + optimizer top-20	0.92	25.3	20.4	1704

Table A2. Robustness of the BigTech consensus configuration under stress. Moderate noise and mean rescaling have limited effect in the archived runs, whereas feature dropout produces the clearest deterioration in Sharpe and turnover.

Stress setting	Sharpe ↑	MDD ↓	Turnover ↓
Base consensus	2.06	16.8	0.081
Noise 0.01	2.00	**16.6**	0.082
Noise 0.02	1.93	**16.6**	0.081
Dropout 0.05	1.96	19.9	0.094
Dropout 0.10	1.80	17.5	0.108
μ scale 0.8×	2.06	16.8	0.081
μ scale 1.2×	2.06	16.8	0.081

 doi:10.20944/preprints202605.0517.v1

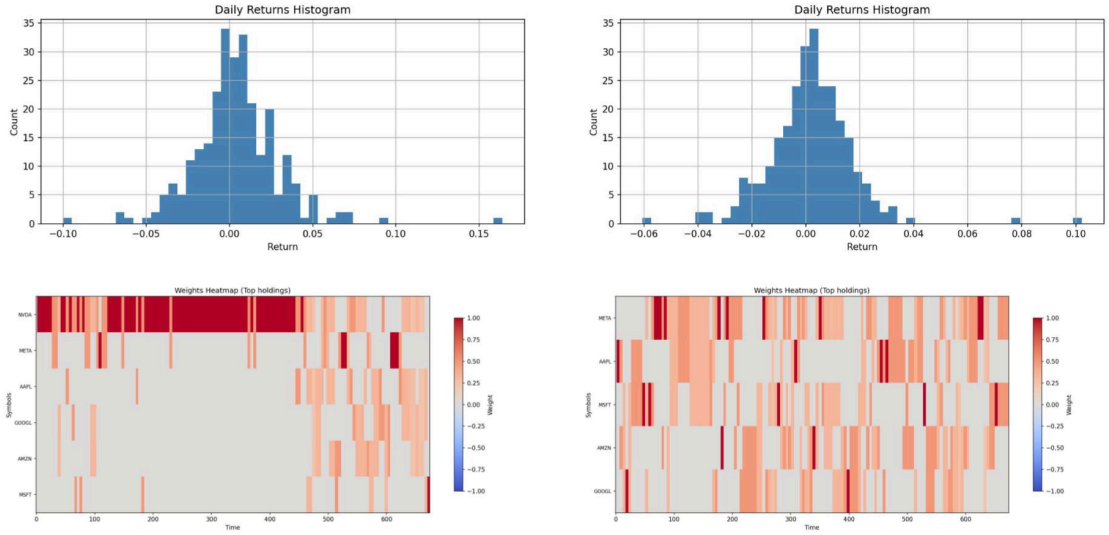

Figure A1. Daily return distributions and representative portfolio-weight heatmaps from the current artifact. The heatmaps show that the hierarchy remains sparse in practice, concentrating exposure in a limited active subset rather than spreading capital uniformly across each universe.

Complexity and implementation notes

Feature construction and forecasting scale roughly linearly with the number of assets, while allocation is dominated by covariance estimation and matrix inversion on the selected universe. The meta-policy adds little overhead, requiring only one validation-quarter evaluation per mode, which can be run in parallel. In practice, each run completes in minutes on a CPU workstation, making the operating-point analyses in Section 5 practical without large-scale training infrastructure.

Appendix C. Reproducibility and Artifact Notes

Runtime profile

The experiments in the current artifact run on a CPU workstation and complete in minutes per configuration. The main computational burden comes from repeated backtesting and covariance-based allocation rather than from large-scale neural training.

Hyperparameter ranges

The principal search ranges used in the archived runs are listed in Table A3. These values are reported to clarify the scale of the sweeps rather than to claim exhaustive tuning.

Table A3. Representative hyperparameter ranges used in the archived experiments.

Component	Hyperparameter	Values / Range
Router	Top-K experts	$\{1, 2, 3, 5\}$
Router	Minimum confidence	$\{0.01, 0.05, 0.1\}$
Router	Signal cost (bps)	$\{0, 5, 10\}$
Allocator	Risk aversion	$\{2, 5, 8\}$
Allocator	Turnover control	enabled / disabled
Universe	Top-n assets	$\{10, 15, 20, 25\}$

Artifact contents

The local artifact includes the Python codebase, backtest summaries, turnover traces, return curves, and experiment-level reports used in the paper. Some directories contain repeated or auxiliary runs; the paper therefore cites only the specific result files used for the main tables. This is intended to make the final narrative traceable to concrete experiment outputs. During revision, we also explored

 doi:10.20944/preprints202605.0517.v1

an additional cross-market path based on merged Chinese equity data stored outside the main U.S. JSON archive. That data source was partially converted into the local backtesting format, but the resulting evaluation window did not yet produce a valid non-empty equity curve under the same 2024–2025 protocol. We therefore do not report those exploratory CN runs in the main paper.

Appendix D. Use of LLM

LLM-based tools were used during the development of this project for limited coding assistance and manuscript editing. All code, analyses, numerical claims, and final text were reviewed and revised by the authors before inclusion. The reported results were taken from the local experimental artifact rather than generated by an LLM.

Appendix E. Ethics Statement

This work studies portfolio decision-making in a high-stakes setting where misuse of backtested systems can cause financial harm. The intended use is methodological research on hierarchical, constraint-aware investment agents; it is not financial advice and is not intended for autonomous live deployment. Experiments use historical market data and derived features only; no human-subject data or personally identifiable information are involved. Key risks include distribution shift, overfitting, concentration, and over-trust in simulated results. We partially mitigate these risks by enforcing turnover and portfolio constraints, reporting downside metrics together with returns, and discussing failure modes such as sensitivity to missing signals. We release the project as a research artifact and do not provide brokerage connectivity or live execution tooling.

Appendix F. Limitations

First, the empirical scope is limited to a bounded set of U.S. equity backtests, so the strongest claims can be interpreted as evidence about operating-point behavior within this artifact rather than as universal claims across markets and regimes. Second, the hierarchy is more sensitive to missing information than to moderate additive noise, as shown by the dropout robustness tests in Section 5. Third, some sweeps in the archived artifact collapse to numerically identical outcomes, which is useful evidence of stable defaults but limits how strongly we can argue for fine-grained router or aggregator effects.

Appendix G. Impact Statement

This work concerns automated support for high-stakes financial decision-making. Its most positive potential impact is methodological: it offers a principled framework for studying when and how multiple financial tools should be coordinated before capital is allocated.

Societal Consequences

Financial risk: Backtest performance can be mistaken for deployable trading quality. We emphasize that the paper is a research benchmark rather than a production system. **Automation risk:** Hierarchical agents may make autonomous portfolio decisions appear more reliable than they are. Human oversight and audit trails remain necessary. **Bias and concentration:** Financial models can inherit biases from historical data. The hierarchy partly mitigates this through portfolio caps and turnover control, but careful monitoring remains essential.

Appendix H. Usage of Large Language Models

The authors used large language models only for polishing prose of text where the complete draft was fully written by the authors initially and polished later with the help of LLM-based assistants including ChatGPT, Gemini, and Perplexity. The authors' used code assistants including Cursor and Copilot to implement the authors' original design and ideas. The scientific contributions, technical methods, ideas and core results are entirely the original work of the authors.

 doi:10.20944/preprints202605.0517.v1

References

1. Ionescu, S.A.; Diaconita, V. Transforming financial decision-making: the interplay of AI, cloud computing and advanced data management technologies. *International Journal of Computers Communications & Control* **2023**, *18*.
2. Gunjan, A.; Bhattacharyya, S. A brief review of portfolio optimization techniques. *Artificial Intelligence Review* **2023**, *56*, 3847–3886.
3. Choi, K.; Hammoudeh, S. Volatility behavior of oil, industrial commodity and stock markets in a regime-switching environment. *Energy policy* **2010**, *38*, 4388–4399.
4. De Prado, M.L. Building diversified portfolios that outperform out-of-sample. *Journal of Portfolio Management* **2016**, *42*, 59–69.
5. Ledoit, O.; Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **2004**, *88*, 365–411.
6. Markowitz, H.M.; Todd, G.P. *Mean-variance analysis in portfolio choice and capital markets*; John Wiley & Sons, 2000.
7. Ahmadi-Javid, A. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications* **2012**, *155*, 1105–1123.
8. Elmachtoub, A.N.; Grigas, P. Smart "predict, then optimize". *Management Science* **2022**, *68*, 9–26.
9. Donti, P.; Amos, B.; Kolter, J.Z. Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems* **2017**, *30*.
10. Wilder, B.; Dilkina, B.; Tambe, M. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 1658–1665.
11. Kou, Z.; Yu, H.; Luo, J.; Peng, J.; Li, X.; Liu, C.; Dai, J.; Chen, L.; Han, S.; Guo, Y. Automate strategy finding with llm in quant investment. *arXiv preprint arXiv:2409.06289* **2024**.
12. Hwang, Y.; Kong, Y.; Zohren, S. Decision-informed neural networks with large language model integration for portfolio optimization. *arXiv preprint arXiv:2502.00828* **2025**.
13. Bo, L.; Liao, H.; Yu, X. Risk sensitive portfolio optimization with default contagion and regime-switching. *SIAM Journal on Control and Optimization* **2019**, *57*, 366–401.
14. Zhang, W.; Zhao, L.; Xia, H.; Sun, S.; Sun, J.; Qin, M.; Li, X.; Zhao, Y.; Zhao, Y.; Cai, X.; et al. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In Proceedings of the Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining, 2024, pp. 4314–4325.
15. Shi, Y.; Fu, Z.; Chen, S.; Zhao, B.; Xu, W.; Zhang, C.; Li, J. Kronos: A foundation model for the language of financial markets. *arXiv preprint arXiv:2508.02739* **2025**.
16. Luo, J.; Kou, Z.; Yang, L.; Luo, X.; Huang, J.; Xiao, Z.; Peng, J.; Liu, C.; Ji, J.; Liu, X.; et al. FinMME: Benchmark Dataset for Financial Multi-Modal Reasoning Evaluation. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 29465–29489. https://doi.org/10.18653/v1/2025.acl-long.1426.
17. Nixon, J.; Dusenberry, M.W.; Zhang, L.; Jerfel, G.; Tran, D. Measuring calibration in deep learning. In Proceedings of the CVPR workshops, 2019, Vol. 2.
18. Team, K.; Bai, T.; Bai, Y.; Bao, Y.; Cai, S.; Cao, Y.; Charles, Y.; Che, H.; Chen, C.; Chen, G.; et al. Kimi K2. 5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276* **2026**.
19. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* **2017**.
20. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **2022**, *23*, 1–39.
21. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* **2024**.
22. Abdulaziz, A.; Zhou, J.; Di Fulvio, A.; Altmann, Y.; McLaughlin, S. Semi-supervised gaussian mixture variational autoencoder for pulse shape discrimination. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 3538–3542.
23. Maire, F.; Badmaev, D.; Braik, F.; Gkartzonika, I.; Polymenakos, L.; Sfakianakis, P.; Tirchas, P. Agentic AI Framework for Technical Excellence: A Discipline-Based, Scalable Multimodal Assistant for Subsurface. In Proceedings of the International Petroleum Technology Conference. IPTC, 2026, p. D021S013R006.

 doi:10.20944/preprints202605.0517.v1

24. Wang, J.; Liu, J.; Fu, Y.; Li, Y.; Wang, X.; Lin, Y.; Yue, Y.; Zhang, L.; Wang, Y.; Wang, K. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265* **2025**.
25. Filos, A. Reinforcement learning for portfolio management. *arXiv preprint arXiv:1909.09571* **2019**.
26. Ye, Y.; Pei, H.; Wang, B.; Chen, P.Y.; Zhu, Y.; Xiao, J.; Li, B. Reinforcement-learning based portfolio management with augmented asset movement prediction states. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 34, pp. 1112–1119.
27. Swaminathan, A.; Joachims, T. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems* **2015**, *28*.
28. Dudík, M.; Langford, J.; Li, L. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601* **2011**.
29. Jiang, N.; Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In Proceedings of the International conference on machine learning. PMLR, 2016, pp. 652–661.
30. Thomas, P.; Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In Proceedings of the International conference on machine learning. PMLR, 2016, pp. 2139–2148.
31. Fakoor, R.; Mueller, J.W.; Asadi, K.; Chaudhari, P.; Smola, A.J. Continuous doubly constrained batch reinforcement learning. *Advances in Neural Information Processing Systems* **2021**, *34*, 11260–11273.
32. Carlstrom, C.T.; Fuerst, T.S. Investment and interest rate policy: a discrete time analysis. *Journal of Economic Theory* **2005**, *123*, 4–20.
33. Kahraman, C.B.; Tookes, H. Leverage constraints and liquidity: What can we learn from margin trading. *Journal of Finance* **2014**.
34. Brunnermeier, M.K.; Eisenbach, T.M.; Sannikov, Y. Macroeconomics with financial frictions: A survey **2012**.
35. Dai, M.; Xu, Z.Q.; Zhou, X.Y. Continuous-time Markowitz's model with transaction costs. *SIAM Journal on Financial Mathematics* **2010**, *1*, 96–125.
36. Chan, L.K.; Jegadeesh, N.; Lakonishok, J. Momentum strategies. *The journal of Finance* **1996**, *51*, 1681–1713.
37. Wilder, J.W. *New concepts in technical trading systems*; Greensboro, NC, 1978.
38. Fang, F.; Ventre, C.; Basios, M.; Kanthan, L.; Martinez-Rego, D.; Wu, F.; Li, L. Cryptocurrency trading: a comprehensive survey. *Blockchain, Crypto Assets, and Financial Innovation: A Decade of Insights and Advances* **2025**, pp. 55–127.
39. Fior, J.; Cagliero, L. A risk-aware approach to stock portfolio allocation based on Deep Q-Networks. In Proceedings of the 2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT). IEEE, 2022, pp. 1–5.
40. Reyes, J.; Di Jorio, L.; Low-Kam, C.; Kersten-Oertel, M. Precision-Weighted Federated Learning. *Computational Intelligence* **2025**, *41*, e70150.
41. Jegadeesh, N.; Titman, S. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance* **1993**, *48*, 65–91.
42. Sharpe, W.F.; et al. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management* **1998**, *3*, 169–85.
43. Lo, A.W. The statistics of Sharpe ratios. *Financial Analysts Journal* **2002**, *58*, 36–52.
44. Uehara, M.; Shi, C.; Kallus, N. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355* **2022**.
45. Dudík, M.; Erhan, D.; Langford, J.; Li, L. Doubly robust policy evaluation and optimization. *Statistical Science* **2014**, *29*, 485–511.
46. Muhle-Karbe, J.; Sefton, J.; Shi, X. Dynamic portfolio choice with intertemporal hedging and transaction costs. *Management Science* **2025**.