Aegis: Uncertainty-Aware Governance for AI-Generated Signals

Yi-Ting, Chiu
National Tsing Hua University
Hsinchu, Taiwan
ericchiu801@gmail.com

Abstract

We introduce **Aegis**, a two-layer governance *framework* that makes AI-generated financial signals risk-aware and auditable. Aegis integrates a pluggable Signal Generator with an Intelligent Gatekeeper that filters signals using model-derived uncertainty quantification (UQ) and a regime proxy. In backtests on Taiwan equities (2003–2025), the framework reduced maximum drawdown from 34% to 4% and improved Sharpe from 1.3 to 1.7. By exploiting the monotonic relation between interval width and error, the Gatekeeper translates UQ into transparent inclusion rules without requiring exact calibration. As a simple, reproducible, and extensible governance framework, Aegis establishes a baseline *verifiable risk agent* for future generative-finance systems.

1 Introduction

Generative AI is increasingly applied in finance, from LLM-driven sentiment factors [1] to synthetic scenario generation. Yet direct use in trading risks overconfidence, regime sensitivity, and weak accountability—concerns already flagged in the NIST AI RMF [2] and FSB guidance [3].

Uncertainty quantification (UQ) is well studied [4, 5], but remains detached from governance. We propose *Aegis*, a simple two-layer framework: a pluggable Signal Generator and an auditable Gatekeeper that filters signals using UQ and regime proxies. In 22-year backtests on Taiwan equities, A cut maximum drawdown from 34% to 4% and raised Sharpe from 1.3 to 1.7.

Our goal is broader than backtest performance: Aegis illustrates a paradigm for *verifiable risk agents*—deterministic modules that embed uncertainty into transparent rules. We argue that such governance layers are essential for trustworthy generative finance.

2 Architecture: Signal Generator + Gatekeeper

2.1 Layer 1: Signal Generator

To isolate the governance effect, we implement Layer 1 as a random forest quantile regressor that produces conditional return quantiles. While quantile levels are configurable, our default outputs are the 75th, 80th, and 85th percentiles, which yield stable uncertainty proxies on noisy financial time series. The design is deliberately modular: the Signal Generator can be replaced with *any* source of signals, including LLM-based sentiment factors or generative scenario models. This ensures that the evaluation of the governance layer is independent of the sophistication of the alpha model, while still establishing a transparent baseline against which more advanced generators can be benchmarked.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

2.2 Layer 2: Intelligent Gatekeeper

The Gatekeeper receives signals from Layer 1 and applies two deterministic inputs:

- **Uncertainty (UQ)** Interval width is defined as $w = q_{0.85} - q_{0.75}$, with larger w indicating lower confidence.
- **Regime** Market regime is proxied by the rolling annualized volatility of a broad market index. A period is classified as "high-vol" if volatility exceeds its expanding-window 80th percentile, computed using past-only data. We test lookback windows of 60, 90, and 120 trading days to ensure robustness.

Decision rule (auditable). In normal regimes, a signal is accepted if its confidence ≥ 0.50; in high-vol regimes, the threshold is raised to 0.75. Thresholds were fixed ex-ante using an anchored training window and then held constant out-of-sample. Although ensemble intervals under-cover in absolute terms (90% nominal $\approx$ 73% empirical), width remains monotonically related to error, making it suitable as a *relative* confidence proxy. Post-hoc conformal recalibration could restore nominal coverage without affecting this ranking. The full procedure is provided in Algorithm 1 (Appendix C).

3 Experimental Setup

3.1 Universe & Horizon

Our dataset consists of 5,579 daily observations spanning January 2, 2003 to August 26, 2025. The universe includes approximately 500 large-cap, liquid equities listed on the Taiwan Stock Exchange, representing the investable subset of the market rather than the full cross-section. Returns are expressed on a total-return basis with delisted securities retained to avoid survivorship bias. Corporate actions and dividend reinvestments are incorporated to construct a bias-free panel suitable for long-horizon evaluation. While our raw source is a standard financial database, equivalent series can be reconstructed from public APIs; we release code and rules so experiments are reproducible using either proprietary or public inputs.

3.2 Rebalancing & Timing

Portfolios are rebalanced monthly, aligned with calendar month-ends. Signals are generated using information up to day t, with trades executed at $t{+}1$ to avoid look-ahead bias. Over the 2003–2025 sample, this results in 268 monthly rebalancing periods. Positions are long-only and unlevered. Results are reported gross of transaction costs, though we include a simple cost-sensitivity module (e.g., 10 bps per side) to confirm robustness.

3.3 Strategies

We benchmark three variants to isolate the effect of governance:

- **Baseline.** Layer 1 (Random Forest quantile regressor) only.
- **UQ Gating.** Signals admitted only if confidence ≥ 0.50.
- **UQ + Regime-Aware.** Threshold raised to 0.75 during high-volatility regimes, following prior work on regime-switching allocation [6].

3.4 Metrics

Evaluation uses annualized Sharpe ratio, maximum drawdown (MDD), and regime-conditioned performance. Uncertainty quality is further assessed via empirical coverage and monotonicity tests (Section 4.2), consistent with the calibration literature [7].

4 Results

4.1 Performance & Risk Mitigation

Figure 1 illustrates stepwise improvements as governance components are added. The left panel shows cumulative returns, while the right panel plots drawdowns. The baseline strategy achieves higher peaks but suffers from deep and prolonged losses. Uncertainty gating substantially reduces tail losses, and combining gating with regime awareness yields the most balanced risk–return profile.

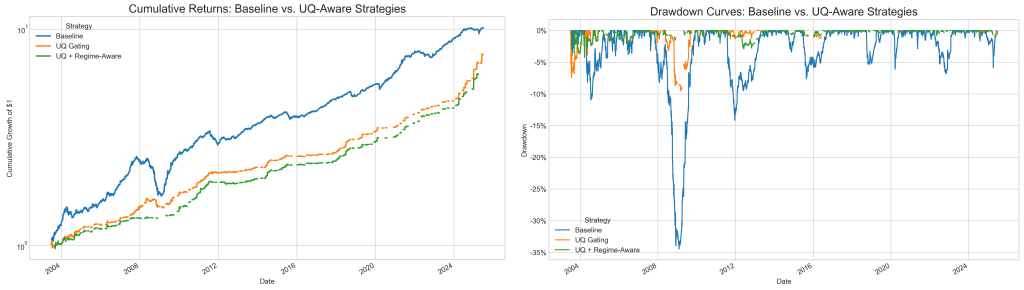

Figure 1: Cumulative returns (left) and drawdowns (right) for baseline, UQ gating, and UQ+Regime-Aware strategies. Governance layers reduce the depth and duration of drawdowns while preserving upside. Flat segments in the gated strategies correspond to periods with no exposure; in practice, idle capital could be parked in risk-free assets (e.g., T-bills), but here we show pure cash to isolate the governance effect.

Table 1 quantifies these patterns: Sharpe improves from 1.30 (baseline) to 1.53 (UQ gating) and 1.68 (UQ+Regime-Aware), while maximum drawdown falls from 34.47% to 9.51% and 4.24%, respectively.

Table 1: Governance layer materially reduces tail risk while preserving returns.

Strategy	Overall Sharpe	Overall MDD	High-Vol Sharpe	High-Vol MDD
Baseline	1.30	34.47%	0.23	34.89%
UQ Gating	1.53	9.51%	1.02	9.51%
UQ + Regime-Aware	**1.68**	**4.24%**	0.76	1.52%

The baseline delivers the highest cumulative return (Figure 1, left) by remaining fully exposed, but this also produces severe drawdowns of more than 34% (Figure 1, right). In contrast, the governance strategies sacrifice some upside to preserve capital, raising Sharpe from 1.30 to 1.68 and cutting drawdowns by nearly an order of magnitude. In practice, where capital continuity and tail-risk control are paramount, such stability is typically valued above raw return maximization.

4.2 Uncertainty Calibration

We validate the uncertainty proxy via two diagnostics:

Coverage. Intervals constructed from our chosen quantiles (75th/80th/85th) under-cover relative to their nominal interpretation: empirical coverage is 72.6% overall, 76.6% in high-vol, and 71.7% in normal regimes—consistent with the over-confidence of uncalibrated ensembles [7].

Monotonicity. Interval width $w = q_{0.85} - q_{0.75}$ is positively related to absolute error (Pearson $= 0.305$, Spearman $= 0.253$). Mean absolute error rises monotonically across deciles, from 3.01% to 11.17% (Figure 2).

Because the gatekeeper relies on the confidence ranking *ordinal* rather than absolute calibration, this monotonic relationship suffices for effective screening. Post-hoc conformal recalibration could restore nominal coverage without altering the ranking.

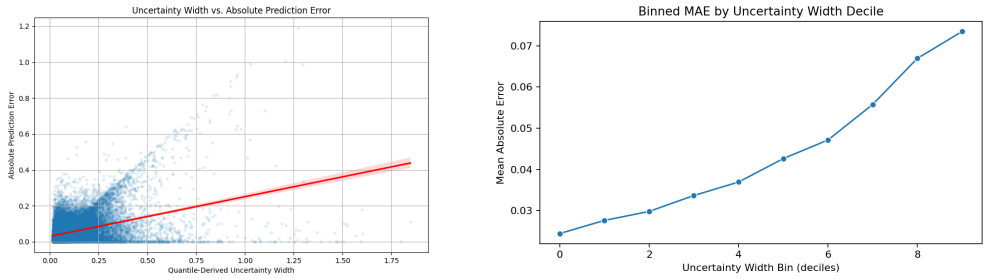

Figure 2: Uncertainty calibration with $w = q_{0.85} - q_{0.75}$. Left: width vs. absolute error (Pearson $= 0.305$, Spearman $= 0.253$). Right: MAE rises monotonically across deciles ($3.01\% \rightarrow 11.17\%$), showing width is a reliable relative confidence proxy.

4.3 Sensitivity

To test robustness, we vary both the volatility lookback window (60, 90, 120) and the Gatekeeper confidence threshold (0.75, 0.80, 0.85). Results are stable: Sharpe ratios consistently exceed 1.6 and maximum drawdowns remain within 4%–6.5% (Table 2).

Table 2: Performance robustness to regime parameters (best-performing threshold per window).

Window	Threshold	Sharpe	MDD
90	0.80	**1.68**	**4.24%**
60	0.75	1.62	6.00%
120	0.85	1.61	6.46%

Full results for all window/threshold combinations are provided in Appendix A, confirming that the conclusions are unchanged across the grid. We also tested a naïve soft-sizing scheme based on interval width, but it produced unstable concentration; robust sizing with width floors and top-K caps is left for future work.

Overall, these results highlight the gatekeeper's role as a verifiable risk agent: a deterministic, auditable rule set that translates model-derived uncertainty into transparent inclusion/exclusion criteria. Such transparency is critical for institutional deployment, where governance frameworks (e.g., NIST AI RMF [2], FSB guidance [3]) require agentic modules to be explainable and subject to audit.

5 Conclusion & Outlook

We have shown that *Aegis*—a simple, auditable governance layer combining uncertainty quantification and regime awareness—substantially improves the risk profile of AI-generated financial signals. The design is reproducible, modular, and serves as a baseline *verifiable risk agent* for generative finance.

Generative integration. While demonstrated with ensemble regressors, the same Gatekeeper can wrap richer sources such as LLM sentiment [1], synthetic scenarios, or foundation models [5].

Scenario generation. The framework extends to regime-conditioned augmentation: forward-looking paths, portfolio CVaR estimates, and exposure scaling when tail risk exceeds preset bounds, linking generative sampling directly to portfolio action.

Toward trustworthy generative finance. Aegis illustrates a principle: generative AI in finance should be valued not just for predictive power, but for its ability to integrate into institutional guardrails [4, 2, 3, 8]. By embedding uncertainty into deterministic, auditable rules, we take a step toward financial AI that is not only profitable, but also *verifiable, accountable, and aligned with systemic safety*.

References

[1] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

[2] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical report, NIST, 2023.

[3] Financial Stability Board. The financial stability implications of artificial intelligence. Technical report, Financial Stability Board, 2024.

[4] Txus Blasco, Vicente García, and J. Salvador Sánchez. A survey of uncertainty quantification in deep learning for financial time series prediction. *Preprint, SSRN*, 2023. Available at SSRN: https://ssrn.com/abstract=4506769.

[5] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25), Vol. 2*, pages 3–7, Toronto, ON, Canada, 2025. ACM.

[6] Yizhan Shu, Chenyu Yu, and John M. Mulvey. Downside risk reduction using regime-switching signals. *arXiv preprint arXiv:2402.05272*, 2024.

[7] Pascal Pernot. About calibration in machine learning uncertainty quantification. *arXiv preprint arXiv:2309.06240*, 2023.

[8] Gillian K. Hadfield and A. Dawson Clark. Regulatory markets: The future of ai governance. *SSRN*, 2023.

A Complete Sensitivity Results

Table 3: Sensitivity results across volatility windows and confidence thresholds.

Window	Threshold	Annual Return	Sharpe	MDD
60	0.75	7.2%	1.62	6.00%
60	0.80	6.5%	1.45	7.20%
60	0.85	6.8%	1.38	7.50%
90	0.75	8.1%	1.40	5.80%
90	0.80	9.4%	**1.68**	**4.24%**
90	0.85	8.7%	1.32	6.20%
120	0.75	7.5%	1.28	6.80%
120	0.80	8.0%	1.34	7.10%
120	0.85	9.0%	1.61	6.46%

Table 3 reports the full grid of volatility windows $(60, 90, 120)$ and Gatekeeper thresholds $(0.75, 0.80, 0.85)$. While the main text highlights the best-performing setting (Table 2), the complete grid shows that Sharpe ratios remain 1.3–1.7 and drawdowns 4%–7.5% across all parameterizations. Annualized returns vary modestly (6%–9%), confirming that governance performance is robust to reasonable changes in regime definition or threshold choice.

B Additional Performance Diagnostics

For completeness, we provide detailed diagnostic plots for each strategy variant. Figures 3–4 report cumulative returns, monthly return distributions, drawdown paths, and boxplots. These diagnostics highlight distinct patterns: the Baseline strategy exhibits fatter left tails and prolonged drawdowns, UQ gating reduces extreme losses at the cost of occasional missed upside, and the UQ+Regime-aware variant delivers the most stable return distribution. This aligns with the Gatekeeper's design objective of controlling downside risk while preserving competitive upside.

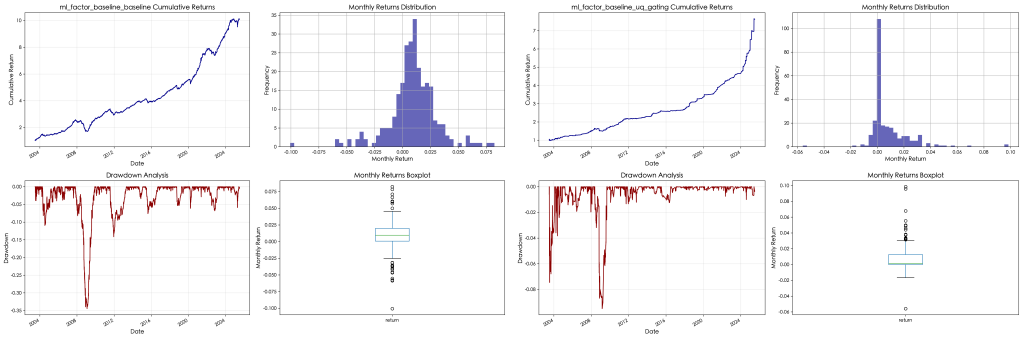

Figure 3: Diagnostics for Baseline (left) and UQ gating (right) strategies. UQ gating reduces left-tail losses and shortens drawdowns while preserving a similar median; axes use different y-ranges.

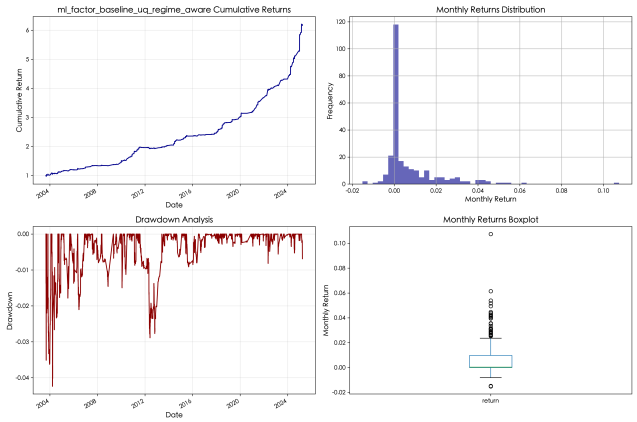

Figure 4: Diagnostics for the UQ+Regime-aware strategy. Regime-aware gating delivers the most stable distribution, with markedly thinner left tails and smoother drawdown paths, confirming effective tail-risk control.

C Algorithmic Specification (Main Method)

For transparency, Algorithm 1 outlines the full procedure of our two-layer governance framework. This is the design used in the main experiments.

Algorithm 1 Two-Layer Governance Framework (Gating Variant)

Require: Equity panel $\mathcal{D}$; lookback window $L \in \{60, 90, 120\}$; thresholds $T \in \{0.75, 0.80, 0.85\}$
Ensure: Portfolio decisions with verifiable gating

```
for each month m do
    Train quantile random forest on D_{≤m} to predict (q_0.75, q_0.80, q_0.85)
    Compute interval width w = q_0.85 − q_0.75; map to confidence score c
    Estimate regime: high-vol if 90-day rolling vol > 80^th percentile (expanding window)
    if regime = normal then
        Accept signal if c ≥ 0.50
    else
        Accept signal if c ≥ T
    end if
    Update portfolio (long-only, equal-weight top K if accepted; else hold cash)
end for
Evaluate Sharpe, MDD, coverage, monotonicity
```

D Experimental Specification

We ran all experiments on an internal research platform. Due to data licensing and platform dependencies, we cannot release production code. However, the procedure is fully specified so that others can replicate results using public data.

- **Data & universe.** $\sim$500 large-cap TWSE equities, 2003–2025, total-return series with delistings retained. Equivalent panels can be reconstructed from public APIs (e.g., exchange open data or commercial providers), with dividend reinvestment.
- **Rebalancing.** Monthly at calendar month-ends. Signals formed using information up to day t; trades executed at $t{+}1$ (no look-ahead). Long-only, unlevered. Results reported gross of costs; a 10 bps/side sensitivity check is included.
- **Signal generator (Layer 1).** Random-forest quantile regressor emitting $(q_{0.75}, q_{0.80}, q_{0.85})$ for next-period returns. Interval width $w = q_{0.85} - q_{0.75}$ serves as a UQ proxy.
- **Regime proxy.** Rolling annualized volatility of a broad market index. A period is "high-vol" if volatility exceeds its expanding-window 80th percentile. We test lookbacks $\{60, 90, 120\}$ trading days.
- **Gatekeeper (Layer 2).** Map width to a confidence score (smaller $w \Rightarrow$ higher confidence). Accept if confidence ≥ 0.50 in normal regimes; raise threshold to $\geq \{0.75, 0.80, 0.85\}$ in high-vol regimes.
- **Idle capital.** When signals are filtered out, capital is held in cash for isolation. In practice, capital could be parked in risk-free assets (e.g., T-bills) without changing conclusions.
- **Evaluation.** Annualized Sharpe ratio, maximum drawdown, and regime-conditioned metrics. UQ diagnostics include empirical coverage and monotonicity (error vs. width).

This specification enables third parties to rebuild the dataset and replicate results without access to the internal platform.

E Python-style Pseudocode

For concreteness, we provide a procedural pseudocode in Python style (Listing E), which mirrors Algorithm 1.

```
# Data prep
PANEL = load_equity_panel(public_APIs, start="2003-01-02", end="2025-08-26")
PANEL = apply_dividends_and_delistings(PANEL)

# Monthly loop
for month in chronology(PANEL):
    X_m, y_m = build_features_and_labels(PANEL, up_to=month.end)

    # Layer 1: Quantile RF
    model = RandomForestQuantileRegressor()
    q75, q80, q85 = model.predict_quantiles(X_m, levels=[0.75, 0.80, 0.85])
    width = q85 - q75
    confidence = normalize_inverse(width)

    # Regime proxy
    vol = annualized_vol(index_returns, lookback=L, up_to=month.end)
    is_high_vol = vol > expanding_percentile(index_vol_history, 0.80)

    # Gatekeeper
    T = 0.75 if is_high_vol else 0.50
    ACCEPT = (confidence >= T)

    # Portfolio update
    if ACCEPT:
        weights = construct_long_only_portfolio(q80)
    else:
        weights = zeros_like_universe() # cash

    portfolio = apply_trades(portfolio, weights, exec_at=month.next_open,
        cost_bps=10)

# Evaluation
report_sharpe_MDD(portfolio)
coverage = empirical_coverage(interval=[q75,q85], realized=y_m)
corr_pearson, corr_spearman = corr(width, abs(y_m - q80))
```

F Sizing Variant (Ablation)

For completeness, we report a soft-sizing variant (Listing F) considered during development. Portfolio weights scale with the prediction and inversely with uncertainty width. As noted in Sectiontion 5, this approach concentrated excessively in a few positions and generated unstable drawdowns without safeguards (e.g., width floors, Top-K caps). We therefore treat it as an ablation and focus the main paper on the auditable gating design.

```
# Layer 1: Quantile RF (q05, q50, q95)
prediction = q50
uncertainty_width = q95 - q05

# Regime-aware threshold
vol = annualized_vol(index_returns, lookback=90, up_to=month.end)
is_high_vol = vol > percentile(index_vol_history, 80)
prediction_threshold = median(prediction[prediction > 0]) if is_high_vol else 0

# Soft sizing
weights = prediction / (uncertainty_width + 1e-6)
weights[prediction < prediction_threshold] = 0
weights = normalize(weights)

portfolio = apply_trades(portfolio, weights, exec_at=month.next_open, cost_bps=10)
```

G Experimental Setup and Computational Resources

G.1 Hardware

All experiments were conducted on a single consumer laptop, without specialized accelerators or clusters.

- **Machine:** MacBook Pro 14" (2021)
- **Chip:** Apple M1 Pro (8-core CPU, 14-core GPU)
- **Memory:** 16 GB RAM
- **Storage:** 1 TB SSD

G.2 Software Environment

Python (3.11) with standard open-source libraries: `pandas` (2.0.3), `numpy` (1.24.3), `scikit-learn` (1.3.0), `matplotlib` (3.7.2). A frozen environment file is provided in the supplementary material.

G.3 Execution Time and Compute Cost

The computational cost is modest:

- A single backtest run (Baseline, UQ Gating, UQ + Regime-Aware, plus sensitivity grid over 9 parameter combinations) completed in **under 15 minutes**.
- Auxiliary diagnostics (uncertainty calibration, ablation) each completed in **2–3 minutes**.

The **total wall-clock time to reproduce all tables and figures in this paper is less than 20 minutes** on the above hardware. No GPUs, distributed training, or large-scale hyperparameter sweeps were required. This low compute footprint makes the framework easily reproducible on commodity hardware.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state a governance-layer contribution (gating + regime-awareness), not a new alpha model; results and scope in Sections. 2, 4 match the abstract's claims.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We note under-coverage of prediction intervals, reliance on ordinal UQ, cash-when-filtered assumption, and deferred robust sizing; see Sections. 4.2, 4.3, 5.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is empirical/methodological without new formal theorems or proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?

   Answer: [Yes]

   Justification: Full procedure, data construction rules, and algorithmic/pseudocode specs are given in Appendix C, D, E; sensitivity grids reported in Section. 4.3 and Appendix A.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce results?

   Answer: [No]

   Justification: Production code and raw data are license-restricted; we provide step-by-step replication specs and public-data reconstruction guidance in Appendix D.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details?

   Answer: [Yes]

   Justification: Universe, horizon, rebalancing, costs, model outputs ($q_{0.75}$, $q_{0.80}$, $q_{0.85}$), regime proxy, thresholds, and evaluation metrics are detailed in Sections. 3, 4.3 and Appendix D.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or significance tests?

   Answer: [No]

   Justification: Time-series backtests emphasize risk-adjusted metrics and robustness via a window/threshold grid; formal CIs are left for future work given monthly rebalancing and dependency structure.

8. **Experiments compute resources**

   Answer: [Yes]

   Justification: All experiments were run on a MacBook Pro 14 (2021, Apple M1 Pro, 16GB RAM). A full reproduction of all tables and figures completes in under 20 minutes. Complete details are given in Appendix G.

9. **Code of ethics**

Question: Does the research conform to the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: No human subjects or sensitive content; financial data used under license; results framed for safe institutional deployment with auditable guardrails (Section. 5).

10. **Broader impacts**

    Question: Does the paper discuss potential positive and negative societal impacts?

    Answer: [Yes]

    Justification: We discuss governance benefits, regulatory alignment, and potential misuse/oversizing risks, with mitigations via verifiable agents and conservative gating (Sectiontion. 5).

11. **Safeguards**

    Question: Are safeguards described for high-risk releases?

    Answer: [NA]

    Justification: We do not release models or datasets; the work proposes a governance mechanism rather than a deployable foundation model or scraped corpus.

12. **Licenses for existing assets**

    Question: Are external assets properly credited and licenses respected?

    Answer: [Yes]

    Justification: Prior work is cited; market data use is license-restricted; we provide instructions for reconstructing equivalent public panels without redistributing proprietary data (Appendix D).

13. **New assets**

    Question: Are new assets introduced and documented?

    Answer: [NA]

    Justification: No new public dataset or code release; we provide procedural documentation instead.

14. **Crowdsourcing and research with human subjects**

    Question: Is crowdsourcing or human-subject research involved?

    Answer: [NA]

    Justification: No human participants or annotation tasks are used.

15. **Institutional review board (IRB) approvals or equivalent**

    Question: Are IRB approvals discussed if applicable?

    Answer: [NA]

    Justification: Not applicable; there are no human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important component of the core methods?

    Answer: [NA]

    Justification: LLMs are discussed as future signal sources; the core method and experiments do not rely on LLMs.