Conformal Trust Control for Learned Sequential Policies under Regime Shift

Jiajun Li
Imperial College London
London, United Kingdom
jiajunli3625@163.com

Ce Guo
Imperial College London
London, United Kingdom
c.guo@imperial.ac.uk

Wayne Luk
Imperial College London
London, United Kingdom
w.luk@imperial.ac.uk

Abstract

A learned sequential policy may fail when its deployment distribution differs from training. We present a two-layer trust-control mechanism that anchors a learned residual to a trusted fallback and regulates deployment through a conformal-risk-control-based trust factor selected from completed outcomes. Zero trust recovers the fallback exactly. We evaluate the mechanism on S&P 500 weekly-option hedging across four held-out regimes: 2018, 2020, 2022, and 2024. In the 2022 stress regime, the controller reduces the aggressive policy's realized breach frequency from 27.3% at full trust to 9.6%, retaining a +22.45 bp mean improvement. This mean is 15.2% above that of the uniform-weight conformal baseline, whose realized breach frequency is 6.7%. These are different realized operating points, so the mean difference does not establish superiority at matched breach frequency. In the four-regime deployment-rule comparison, the controller achieves the highest mean among rules with observed breach frequency at or below 10% only in 2022. Across all twelve regime–risk-aversion cells, episode-weighted guarded breach frequencies range from 0.078 to 0.102. One-sided 95% bootstrap upper bounds for date-averaged breach rates exceed 0.10 in every cell. Under a common 5 bp budget, aggressive training followed by attenuation has a higher held-out mean in all twelve cells, but with a substantially heavier lower tail.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Economics*.

Keywords

conformal risk control, deployment reliability, risk management, sequential decision making, regime shift

ACM Reference Format:
Jiajun Li, Ce Guo, and Wayne Luk. 2026. Conformal Trust Control for Learned Sequential Policies under Regime Shift. In *7th ACM International Conference on AI in Finance (ICAIF '26), November 14–17, 2026, Milan, AA, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3858063.3858769

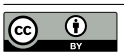
1 Introduction

Financial decision policies are deployed in markets whose volatility, liquidity, and price dynamics change across persistent regimes. However, most learned policies are trained under the implicit assumption that future observations will resemble the historical data on which they were optimized. In financial time series, this assumption is especially fragile, because observations are serially dependent and the underlying process may move between persistent regimes [11]. Therefore, a policy may become least reliable during market stress, when benchmark-relative losses matter most.

Two strands of literature speak to this situation, but neither answers the operational question on its own. Change-point and concept-drift methods identify when the data-generating process has changed [8, 11], while learned hedging focuses on training stronger policies [6, 15]. However, deployment also requires deciding what to execute next. We address this with a deployment-time controller building on conformal risk control [3]. The controller uses realized benchmark-relative outcomes to regulate exposure to a frozen policy.[1] Completed-episode breach feedback updates the working error level and thereby adjusts the trust factor for subsequent episodes (Section 4). This connects observed deployment performance to policy attenuation under regime shift.

Two concrete drawbacks motivate this. First, deploying a learned policy at full strength because it performed well in training is unsafe under shift. In our 2022 stress cell, the aggressively trained residual breaches the deployment budget in 27.3% of episodes, against a 10% target. Applying our trust controller to the same frozen policy reduces the realized breach frequency to 9.6% while retaining a mean improvement of +22.45 bp relative to the fallback (Table 3). Second, identifying regime change [8, 11] does not by itself establish which learned policy to deploy. In the diagnostic comparison of Section 6.2, the within-year oracle has a higher mean than the size-matched historical policy across all three κ values only in 2020, and underperforms it by up to 23.10 bp in 2024. This motivates regulating trust from realized benchmark-relative outcomes rather than regime labels.

Long-call hedging provides a testbed, with the Black–Scholes delta hedge as a training-independent fallback [5]. A breach occurs when deployment underperforms this fallback by more than a prescribed budget. Layer 1 learns a clipped residual policy using a benchmark-relative mean–CVaR objective. Layer 2 selects the largest trust factor λ whose conformal upper loss quantile, computed from completed episodes, is within budget. Setting $\lambda = 0$ recovers the fallback exactly. Trust thus responds to realized deployment outcomes without retraining or assigning regime labels. Hedging provides dated regime shifts, monetary losses, and a strong fallback. Section 3.3 gives four requirements for transfer beyond hedging.



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICAIF '26, Milan, AA, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3087-0/2026/11
https://doi.org/10.1145/3858063.3858769

[1] Code and artifact manifests: https://github.com/sonnets-project/conformal-trust-control

We evaluate residual performance relative to the fallback under distribution shift (Q1), regime-matching versus sample-size effects (Q2), guardrail costs for a stable policy (Q3), aggressive-policy breach frequencies relative to the target (Q4), and common-budget mean differences between aggressive and validation-selected policies (Q5).

This paper makes three contributions. First, it introduces a residual policy anchored to a fallback for managing regime shift, with conformal episode-level trust selected from completed outcomes and exact recovery of the fallback at $\lambda = 0$. Second, it identifies four structural requirements for transfer beyond hedging. Third, a chronological four-regime SPXW evaluation shows low guardrail cost for a stable policy, empirical target tracking under stress, and a higher common-budget mean for aggressive training at the cost of a heavier lower tail.

2 Related Work

Learned hedging. Deep hedging replaces closed-form hedge ratios with neural policies trained on simulated paths [6], typically evaluated with training and test paths from the same stochastic-volatility model [12]. Such designs support controlled comparison but give little evidence about behavior when the deployment regime differs from the training regime. Work on real data [15] and on multi-instrument hedging with distributional critics [7] moves closer to practice through empirical evaluation and risk-sensitive policy learning. Our Layer 1 instead learns a residual correction to the classical hedge rather than the hedge itself.

Safe deployment of learned policies. Safety can be imposed while a policy is trained or supervised once it is fixed. Constrained policy optimization takes the first route, folding cost constraints into learning [1]. The constraint and the policy are then interdependent, so changing the operational budget or the failure definition means optimizing again, whereas our policy is frozen before the deployment rule is calibrated. Shielding monitors proposed actions against a specification and overrides violations before execution [2, 10], which requires the unsafe event to be a state-wise predicate checkable before acting. A hedging loss is path dependent and observed only after prices move, so the event we control admits no such predicate. Runtime assurance architectures, which pair an untrusted high-performance controller with a baseline controller and a switching rule [16, 17], are the closest architectural neighbors of Layer 2. We differ in the nature of the claim: the fallback is an interpretable model-based hedge rather than a verified controller, the monitored failure is a realized episode-level outcome rather than a predicted state violation, and intervention is a graded trust factor rather than a binary switch.

Conformal risk and policy control. Conformal risk control (CRC) extends conformal prediction from coverage to general monotone risks [3], while Conformal Decision Theory calibrates downstream decisions directly [14]. Most closely related, Conformal Policy Control (CPC) regulates an optimized policy relative to a safe reference policy [18]. Our setting differs in the information available at deployment. After an episode resolves, its realized market path can be replayed under every candidate trust factor, directly producing benchmark-relative monetary outcomes without retraining the frozen policy. Therefore, Layer 2 operates on a completed-outcome ledger and adapts trust under chronological market regime shift rather than relying on a pre-assigned regime label.

Nonexchangeability and adaptation. Weighted conformal prediction reweights calibration observations under distribution shift [4], while adaptive conformal inference updates the working error level from realized errors [9]. Our Layer 2 combines these ideas by favoring recent, IV-similar completed episodes and updating the working error level only after outcomes resolve.

Risk-sensitive objectives. CVaR is the standard coherent tail measure, and the Rockafellar–Uryasev variational representation supports gradient-based optimization [19]. We optimize it on paired differences against the fallback rather than on absolute returns, which aligns the training signal with the quantity the deployment layer monitors.

3 Problem Setup

3.1 Episodes and benchmark-relative outcomes

The unit of interaction is a hedging episode indexed by i, with decision times $t = 1, \ldots, T$. Each episode tracks the marked-to-market P&L of a long European call written on an underlying index and hedged through a signed position a_t in the underlying. Here, $a_t < 0$ denotes a short hedge position.

The option's initial premium is common to the paired strategies and cancels from benchmark-relative comparisons. The position is adjusted once per trading day for $t = 1, \ldots, T$. Daily P&L is

$$\mathrm{pnl}_t = a_t(S_{t+1} - S_t) + (C_{t+1} - C_t) - \tau|a_t - a_{t-1}|S_t, \qquad (1)$$

where S_t is the index level, C_t the option price, a_t the hedge position held over day t, and τ the proportional transaction-cost rate. Episode-level P&L and the loss budget are expressed on a common scale. The normalization used in our experiments is specified in Section 5.

Two properties matter later. The horizon is short, so realized outcomes enter the deployment evidence set with small delay. And Eq.(1) is differentiable in the actions except at the origin of the transaction-cost term. Thus, an episode can be differentiated end to end.

3.2 Fallback and benchmark-relative outcome

The fallback is the signed Black–Scholes hedge $a_t^{\mathrm{BS}} = -\Phi(d_1)$, where Φ is the standard normal cumulative distribution function and d_1 is computed from moneyness, the risk-free rate, the contract's implied volatility (IV), and time to maturity. Every strategy is evaluated in a paired design in which the strategy and the benchmark hedge the same contract over the same days under the same cost model. We define the benchmark-relative episode outcome as

$$\Delta\mathrm{PnL}_i(\lambda) = \mathrm{PnL}_i^{\mathrm{deploy}}(\lambda) - \mathrm{PnL}_i^{\mathrm{fallback}},$$

where both P&L terms are net of transaction costs and use the same normalization. This monetary performance measure is distinct from the downside-shaped reward used for training. The two strategies are evaluated on the same contract and realized market path, so the paired difference reflects the effect of changing the hedge rule under a common market realization.

3.3 Structural requirements and scope

The mechanism is not specific to option hedging, but its transfer to another sequential decision problem requires four structural properties.

(1) **An always-admissible fallback.** The fallback must provide an operationally acceptable action whenever the learned component is not used. Since $\lambda = 0$ recovers the fallback, this condition ensures that the candidate set always contains a feasible deployment choice. Without such a fallback, rejecting all positive trust factors would leave the controller with no admissible action to execute.
(2) **Outcomes that resolve quickly relative to the drift.** Layer 2 relies only on completed episodes, so their outcomes must become available before the deployment environment changes substantially. If the outcome delay is comparable to or longer than the persistence of a regime, the ledger will contain stale evidence and the feedback update will react only after the relevant conditions have passed.
(3) **A scalar completed-outcome loss.** Each completed episode must produce an ordered scalar score $L_i(\lambda)$, with a breach defined by $L_i(\lambda) > c$. This scalar ordering is required to form the conformal quantile used by the selection rule. A problem with several incomparable safety criteria would first require an explicit scalarization or a different multi-constraint construction.
(4) **Counterfactual replay.** For every completed episode, the system must be able to evaluate the outcome under each candidate trust factor without retraining the policy. Without replay, the ledger would contain only the outcome at the trust factor that was actually deployed. Comparing all candidates would then require off-policy estimation or active exploration, neither of which the mechanism supports.

The option-hedging setting satisfies all four properties: the Black–Scholes hedge provides the fallback, episode outcomes are observed over a short horizon, benchmark-relative P&L defines a scalar loss, and completed market paths can be replayed at every candidate trust factor. Other applications, such as inventory control relative to a newsvendor baseline or routing relative to a shortest-path heuristic, may satisfy the same requirements when outcomes are observed promptly and counterfactual actions can be evaluated. Settings with long outcome delays, unavailable counterfactuals, or no admissible fallback fall outside the scope of the construction.

4 Method

Figure 1 summarizes the two layers. Layer 1 is trained once and frozen. Layer 2 selects episode-level trust from completed outcomes, so the guardrail can change without policy retraining.

4.1 Layer 1: anchored residual policy

The full-trust Layer-1 policy is

$$a_t = \mathrm{clip}_{\mathcal{A}}\left(a_t^{\mathrm{BS}} + g_\theta(s_t)\right), \tag{2}$$

where $\mathcal{A}$ is the admissible action set containing the fallback action at each decision, $s_t \in \mathcal{S} \subseteq \mathbb{R}^d$ is a real-valued state vector constructed from information available before the decision at time t, and $g_\theta : \mathcal{S} \to \mathbb{R}$ maps the state to a scalar residual adjustment to the fallback hedge. The residual is expressed in the same units as the hedge position, so the quantity inside $\mathrm{clip}_{\mathcal{A}}$ is the candidate action executed by the policy. We require g_θ to support gradient-based end-to-end training, but the construction does not otherwise depend on a particular model architecture or state representation. The action bounds, network architecture, and exact state variables used in our experiments are specified in Section 5.

For Layer 1, let

$$r_t^{(\kappa)} = \mathrm{pnl}_t - \kappa \max(0, -\mathrm{pnl}_t), \qquad \kappa \geq 0,$$

denote a downside-shaped reward. Positive P&L is unchanged, whereas negative P&L is multiplied by $1 + \kappa$, so κ controls the policy's degree of downside aversion. Let $\Delta_i(\theta)$ denote the paired difference between the learned policy's episode reward and that of the fallback on episode i. Layer 1 maximizes

$$\mathcal{J}(\theta) = \mathbb{E}[\Delta_i(\theta)] + \beta\, \mathrm{CVaR}_\alpha(\Delta_i(\theta)), \qquad \alpha \in (0,1),\ \beta \geq 0, \tag{3}$$

where β controls the weight placed on the lower tail of the benchmark-relative episode outcome.

Using the Rockafellar–Uryasev representation [19], the corresponding empirical objective is

$$\widetilde{\mathcal{J}}(\theta, \eta) = \frac{1}{N}\sum_{i=1}^{N} \Delta_i(\theta) + \beta\left(\eta - \frac{1}{\alpha N}\sum_{i=1}^{N}\left(\eta - \Delta_i(\theta)\right)_+\right), \tag{4}$$

which is maximized jointly over θ and the auxiliary scalar η. In our differentiable short-horizon setting, the episode objective can be differentiated directly through the action sequence, so Layer 1 requires neither a critic nor Bellman bootstrapping.

Layer 1 optimizes benchmark-relative performance on the training distribution. Deployment-time trust control is instead the role of Layer 2.

4.2 Layer 2: conformal deployment guardrail

At the start of episode i, Layer 2 selects an episode-level trust factor λ_i from a finite candidate set $\Lambda \subseteq [0,1]$ with $0 \in \Lambda$. The selected value remains fixed throughout the episode, and the deployed action at decision time t is

$$a_{i,t}^{\mathrm{deploy}} = \mathrm{clip}_{\mathcal{A}}\left(a_{i,t}^{\mathrm{BS}} + \lambda_i g_\theta(s_{i,t})\right). \tag{5}$$

Thus, $\lambda_i = 1$ deploys the full learned residual, whereas $\lambda_i = 0$ recovers the fallback exactly. The candidate grid used in the experiments is specified in Section 5.

Breach and budget. Let $\Delta\mathrm{PnL}_i(\lambda)$ denote the benchmark-relative episode outcome obtained at trust factor λ. For a prescribed budget $c > 0$, define the breach indicator

$$B_i(\lambda) = \mathbf{1}\left\{\Delta\mathrm{PnL}_i(\lambda) < -c\right\}. \tag{6}$$

The deployment target is to keep the breach frequency near a prescribed level $\varepsilon \in (0,1)$. This target concerns how often the deployed policy underperforms the fallback by more than c. It does not constrain the magnitude of the loss conditional on a breach. That distinction is examined in Section 6.5.

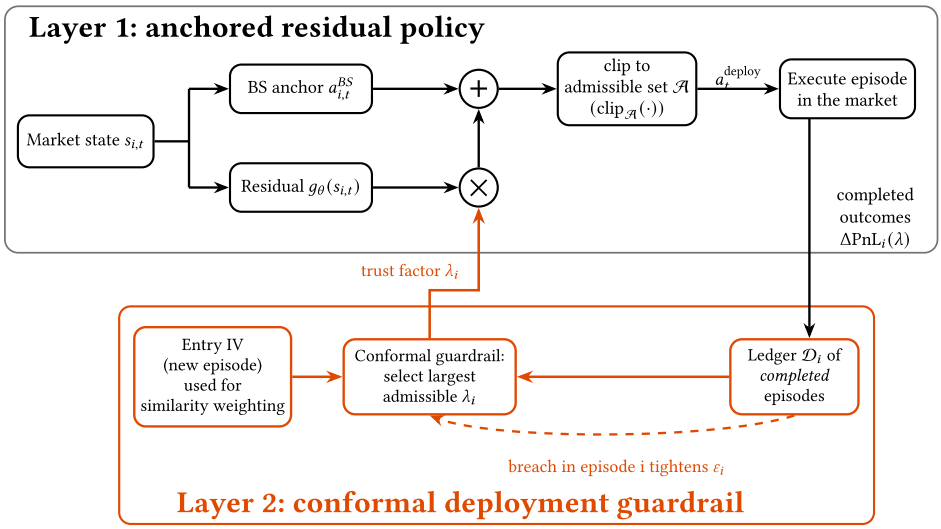


Figure 1: Two-layer trust control, with Layer 1 in black and Layer 2 in orange. Layer 1 adds the trust-scaled neural residual $\lambda_i g_\theta(s_{i,t})$ to the Black–Scholes anchor and clips the resulting action to $\mathcal{A}$. Layer 2 selects the largest admissible trust factor from the completed-outcome ledger $\mathcal{D}_i$. The dashed path shows the feedback loop: completed episode outcomes are scored against the fallback and used to update the working error level ε_i, which governs trust selection for subsequent episodes. Setting $\lambda_i = 0$ recovers the fallback exactly.

Evidence ledger. Before episode i begins, the guardrail may use only the ledger $\mathcal{D}_i$ of episodes whose terminal outcomes have already become available. For every completed episode $j \in \mathcal{D}_i$ and every candidate $\lambda \in \Lambda$, the ledger stores the counterfactual outcome $\Delta\mathrm{PnL}_j(\lambda)$ obtained by replaying that episode at the corresponding trust factor. Replay uses the same frozen policy, realized path, and cost model, and requires no retraining. Running episodes are excluded because their terminal outcomes are unavailable when trust is selected.

Selection rule. Define the benchmark-relative loss score

$$L_j(\lambda) = -\Delta\mathrm{PnL}_j(\lambda). \tag{7}$$

Layer 2 selects the largest admissible positive trust factor, with zero trust as an unconditional fallback:

$$\lambda_i = \max\left(\{0\} \cup \left\{\lambda \in \Lambda \setminus \{0\} : \widehat{Q}^i_{1-\varepsilon_i}\left(\{L_j(\lambda)\}_{j\in\mathcal{D}_i}\right) \le c\right\}\right). \tag{8}$$

Since $\Delta\mathrm{PnL}_j(0) = 0$, zero trust satisfies the benchmark-relative budget by construction and is included independently of the quantile test. If no positive candidate is admitted, Eq. (8) returns $\lambda_i = 0$. We do not assume that $L_j(\lambda)$ is monotone in λ, since position clipping and transaction costs can produce non-monotone outcomes. Each grid candidate is evaluated separately. Selecting the largest admissible λ expresses a preference for retaining residual exposure. This neither guarantees optimal returns nor inherits the standard monotone-CRC guarantee.

For n uniformly weighted calibration scores, we use the k-th smallest loss score, where

$$k = \lfloor (n+1)(1-\varepsilon_i) \rfloor + 1. \tag{9}$$

If $k > n$, the threshold is $+\infty$, so no positive trust factor is admitted. Compared with the standard ceiling rule, this convention uses the next rank when $(n+1)(1-\varepsilon_i)$ is an integer. The ranks agree otherwise. The $n+1$ term reserves probability mass for the unseen episode [20].

Weighting and adaptive feedback. To account for nonstationarity, calibration episodes may receive non-negative relevance weights

$$w_{j,i} = \omega(z_j, z_i; \phi), \tag{10}$$

where z_i contains context available before episode i begins and ϕ denotes fixed weighting parameters. For each candidate λ, the historical loss scores are augmented by a unit-weight score at $+\infty$, and all weights are normalized jointly. After sorting the scores in ascending order, the quantile in Eq. (8) is the first score whose cumulative normalized weight strictly exceeds $1-\varepsilon_i$. If no finite score satisfies this condition, the quantile is $+\infty$. The exact context variables and weighting function used in the option-hedging experiments are specified in Section 5.

After a deployed episode completes, its realized breach indicator is used to update the working error level:

$$\varepsilon_{i+1} = \mathrm{clip}_{[\varepsilon_{\min},\varepsilon_{\max}]}\left(\varepsilon_i + \gamma\left(\varepsilon - B_i(\lambda_i)\right)\right), \tag{11}$$

where $\gamma > 0$ is the adaptation rate. A breach decreases the working error level, making subsequent admission more conservative. By contrast, a non-breach increases the working error level and allows more permissive admission. The update is applied only after the episode outcome becomes observable.

Delayed and concurrent outcomes. Let d index feedback updates. Before trust selection at each decision point, let $\mathcal{B}_d$ contain deployed episodes whose terminal outcomes are available but have not yet been used in a feedback update. With ε_d denoting the working error

level before processing this batch, update

$$\varepsilon_{d+1} = \mathrm{clip}_{[\varepsilon_{\min},\varepsilon_{\max}]}\left(\varepsilon_d + \gamma \sum_{j\in\mathcal{B}_d}\left[\varepsilon - B_j(\lambda_j)\right]\right), \tag{12}$$

where $B_j(\lambda_j)$ uses the trust factor actually deployed on episode j. The updated level ε_{d+1} supplies ε_i for episodes starting at that decision point. No episode contributes to feedback more than once, although its scores may remain in the calibration ledger. An empty batch leaves the working error level unchanged. Summing the increments and clipping once makes the update invariant to the processing order within the batch. Running episodes enter neither the ledger nor the feedback update.

4.3 Reference conformal statement and deployed interpretation

For a trust factor λ and error level ε fixed independently of the calibration data, suppose the calibration and new-episode loss scores are exchangeable and the calibration scores are uniformly weighted. Let $\widehat{q}_{1-\varepsilon}(\lambda)$ be the conservative quantile from Eq. (9), using $\varepsilon_i = \varepsilon$ and setting the quantile to $+\infty$ when $k > n$. This threshold is no smaller than the standard split-conformal threshold with rank $\lceil (n+1)(1-\varepsilon)\rceil$. The usual rank bound [20] therefore gives

$$\Pr\{L_{\mathrm{new}}(\lambda) > \widehat{q}_{1-\varepsilon}(\lambda)\} \le \varepsilon.$$

This marginal guarantee concerns the calibrated quantile. It does not by itself establish a fixed-budget guarantee after data-dependent selection of λ. Our deployed rule additionally uses relevance weights, adaptive feedback, and temporally dependent episodes. We therefore use held-out episode-weighted breach frequencies and bootstrap upper bounds for date-averaged breach rates to assess empirical target tracking. They do not provide population-level compliance certificates.

5 Experimental Setup

Data. We use S&P 500 weekly (SPXW) call options from OptionMetrics IvyDB. Filtering removes non-positive or crossed quotes, relative bid-ask spreads above 50% of mid, missing or non-positive implied volatilities, and prices violating the no-arbitrage intrinsic bound. As a result, roughly 25 million raw daily quote rows are reduced to approximately 5.07 million rows in the cleaned contract-level extract. We keep contracts with 10 to 90 calendar days to maturity and spot-to-strike ratio $S/K \in [0.83, 1.17]$. Index levels come from public market data and the risk-free rate from the one-year Treasury yield.

The cleaned input extract spans 2007–2024. After constructing the five-day episodes and applying the chronological protocol, the analytical walk-forward sample used here spans 2008–2024. The four splits are: train 2008–2016, validation 2017, test 2018; train 2008–2018, validation 2019, test 2020; train 2008–2020, validation 2021, test 2022; and train 2008–2022, validation 2023, test 2024.

Contract coverage is sparse in 2008–2011 and broadens after 2014. We do not impute missing grid cells, so early years contribute fewer training episodes. This affects sample composition but not chronological separation, and all four test years have substantially denser coverage.

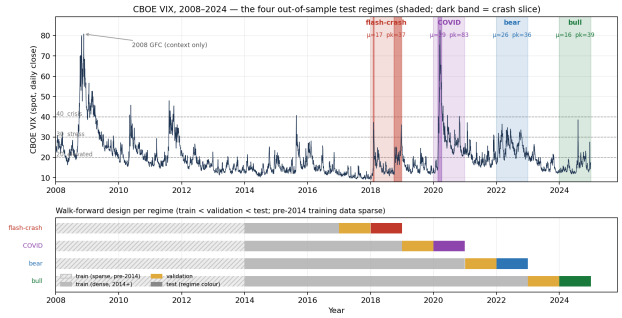


Figure 2: VIX and chronological train–validation–test splits for the primary evaluation. Policy training and model selection use only pre-test data.

Episodes. Start dates are spaced five trading days apart. On each start date we open up to fifteen contracts on the grid $K/S \in \{0.85, 0.925, 1.00, 1.075, 1.15\}$ by DTM $\in \{10, 30, 60\}$ calendar days, selecting for each cell the admissible contract closest to the target. The $K/S = 0.85$ target ($S/K \approx 1.176$) lies marginally outside the moneyness filter, so its contracts are selected at the band edge ($K/S \ge 1/1.17$). Transaction costs are $\tau = 10$ bp per adjustment. The realized test samples contain 602, 663, 675, and 630 episodes in 2018, 2020, 2022, and 2024.

Action constraints. We set $\mathcal{A} = [-1.5, 1.5]$ for the signed underlying position in both Layer 1 training and Layer 2 deployment. The combined position is clipped after adding the residual to the Black–Scholes anchor, with the residual first scaled by λ at deployment. The raw residual is not clipped separately. Since $a_t^{\mathrm{BS}} \in [-1, 0] \subset \mathcal{A}$, setting $\lambda = 0$ recovers the fallback exactly.

Training and model selection. The Layer-1 state contains S_t/K, $T_t/30$, the previous Black–Scholes position, IV, and Black–Scholes delta, gamma, vega, and theta. Greek features use the initial underlying as the episode price scale, with no further standardization. The residual is an 8–64–64–1 ReLU MLP with linear output, trained by full-batch Adam (10^{-3}) for 400 epochs: 100 mean-only, 100 linearly ramping to the target β, and 200 at that target.

We train separate Layer-1 policies for $\kappa \in \{1, 2, 3\}$ with $\alpha = 5\%$ and evaluate $\beta \in \{0, 0.5, 1, 2, 4, 8, 16\}$ on the validation year. Let m_β and q_β denote the validation mean and $\mathrm{CVaR}_{5\%}$ of the paired episode-reward difference. If any $q_\beta > 0$, we select the feasible candidate with the largest m_β. Otherwise, we select the largest-m_β candidate satisfying $q_\beta \ge q_{\max} - 0.25|q_{\max}|$, where $q_{\max} = \max_\beta q_\beta$. The second branch was used in all twelve searches.

Guardrail implementation. In the option-hedging experiments, the candidate trust grid is

$$\Lambda = \{0, 0.25, 0.5, 0.75, 1\},$$

and the nominal breach target is $\varepsilon = 0.10$. Similarity weighting uses each contract's entry implied volatility together with episode age. For a completed episode j and a new episode i, the weight is

$$w_{j,i} = 2^{-\mathrm{age}(j,i)/30} \exp\left(-\frac{|\mathrm{IV}_j - \mathrm{IV}_i|}{h_{\mathrm{IV}}}\right), \qquad h_{\mathrm{IV}} = 0.04, \tag{13}$$

where age(j, i) is the number of calendar days from the completion of historical episode j to the start of episode i. The resulting 30-day half-life and IV bandwidth are fixed a priori and shared across all regimes and all values of κ.

The ACI update uses

$$\gamma = 0.02, \qquad \varepsilon_0 = 0.10, \qquad [\varepsilon_{\min}, \varepsilon_{\max}] = [0.025, 0.40].$$

We refer to the resulting reweighted conformal rule with adaptive feedback as `crc_rwc_aci`.

Protocol. The primary policy evaluation and all Layer 2 experiments use chronological train–validation–test splits. Policy parameters, training hyperparameters, and deployment budgets are fixed before testing, while the guardrail specification is shared across regimes. The oracle comparison in Section 6.2 is a separate, non-deployable diagnostic using part of the test year for training; it is not used for model selection or guardrail calibration. The four test years represent distinct environments: abrupt stress in 2018, the COVID dislocation in 2020, persistent elevation in 2022, and a comparatively calm rising market in 2024. Figure 2 is descriptive only: VIX is neither a policy input nor a traded instrument.

Budget calibration. For each policy, regime–κ cell, and seed, the loss budget is fixed from validation data as

$$c = \max\{5, -\widehat{Q}_{0.10}(\Delta\mathrm{PnL}^{\mathrm{val}}(\lambda = 1))\} \text{ bp}.$$

Thus c is the 10% lower-tail loss of the frozen full-trust policy, subject to a prespecified 5 bp floor, and is fixed before testing. Table entries average the five seed-specific budgets. The common-budget experiment instead fixes $c = 5$ bp for both arms.

Metrics. The primary metric is paired episode $\Delta\mathrm{PnL}_i$, net of transaction costs, expressed in basis points (bp) of the episode's initial underlying index level: the net episode P&L difference is divided by that level and multiplied by 10^4. All reported P&L statistics and deployment loss budgets c use this scale. Across five separately deployed seeds, mean, win rate, and $\mathrm{CVaR}_{5\%}$ are computed from the seed-averaged per-episode path. Pointwise 95% CIs use 3,000 moving-block bootstrap resamples over start-date batches, clustering same-date contracts. The main block length is 10 batches, with 5 and 20 as robustness checks. Breach calculations include all test episodes, including Black–Scholes warm-up. Point estimates give equal weight to episodes within each seed and are averaged across the five seeds. We also report the worst-seed frequency. For the one-sided 95% upper bounds, we first average breach indicators across seeds and compute their mean within each start-date batch. We then give these batch means equal weight in a moving-block bootstrap with 3,000 resamples and blocks of 10 start-date batches. The upper bound is the 95th percentile of the bootstrap means and concerns the date-averaged breach rate.

Baselines. Beyond Black–Scholes, we compare against two published cost-aware hedges: the Hull–White minimum-variance delta [13] and the Whalley–Wilmott no-trade band [21]. For deployment, we compare `crc_rwc_aci` with five published-method-inspired baselines on the same ledger, protocol, and budget. `fix1` (Full) deploys the learned policy at full strength, following the standard deep-hedging convention [6]. `crc` (CRC) uses uniform-weight split-conformal risk control [3, 20]. `crc_rwc` (W-CRC) adds covariate-shift reweighting [4]. `aci_open` (ACI) applies adaptive conformal inference [9] with feedback evaluated at $\lambda = 1$, while `ivrule` (IV) is a detect-then-act rule based on the entry-IV percentile, motivated by regime and drift methods [8, 11]. Our rule uniquely combines context-weighted calibration with closed-loop feedback on realized breaches at the deployed trust factor. Section 6.3 compares these baselines with our complete rule across all four regimes at $\kappa = 2$.

Reproducibility. Every guardrail decision is a weighted order statistic over stored episode outcomes and is deterministic given the ledger, so each trust-factor selection can be replayed exactly and traced to the historical episodes that supported or refused it. Results average five seeds under a pinned single-threaded environment, and an artifact manifest records content hashes of every input table. Code and manifests are available in the repository given in Section 1.

6 Results

We organize the results around the five questions stated in Section 1. Positive paired mean differences favor the learned policy.

6.1 Q1: the anchored residual out of distribution

Takeaway. At full trust, the anchored residual improves mean performance relative to Black–Scholes in all twelve regime–κ cells. In the classical-baseline comparison at $\kappa = 2$, it does not consistently outperform the cost-aware hedges: neither paired mean difference is statistically significant in 2018 or 2020, and Whalley–Wilmott outperforms the residual in 2024.

Figure 3 reports the paired ΔPnL of the full-trust Layer-1 policy relative to Black–Scholes. The estimated mean improvement is positive in all twelve cells, and every pointwise 95% bootstrap interval excludes zero. Win rates range from 64% to 72%. The largest gains occur under stress, reaching +8.90 bp per episode in 2020 and +5.75 bp in 2022 at $\kappa = 1$.

The benchmark-relative $\mathrm{CVaR}_{5\%}$ nevertheless remains negative in every cell. Layer 1 therefore provides a persistent average advantage over Black–Scholes, but its full-trust lower tail does not justify unconditional deployment.

At $\kappa = 2$, Table 1 compares the full-trust, validation-selected residual policy with Whalley–Wilmott [21] and Hull–White [13]. The residual contrasts with Hull–White and Whalley–Wilmott are not statistically distinguishable from zero in 2018 or 2020. In 2022, the residual improves on Hull–White by +3.95 bp [2.46, 5.46] and on Whalley–Wilmott by +7.14 bp [0.42, 14.13]. Both intervals exclude zero, so the improvements are statistically significant under our clustered bootstrap. In the comparatively calm 2024 regime, it is indistinguishable from Hull–White but underperforms Whalley–Wilmott by −2.32 bp [−4.35, −0.48]. Relative performance is therefore not monotone in market stress: the severe 2020 dislocation does not produce a significant residual advantage, whereas Whalley–Wilmott is strongest in calm 2024. This heterogeneity motivates deployment control from realized benchmark-relative evidence rather than a fixed market-stress threshold.

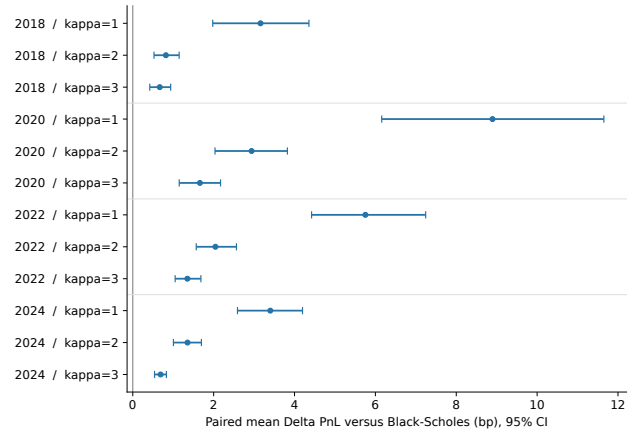


Figure 3: Layer-1 paired ΔPnL against Black–Scholes with 95% bootstrap intervals.

Table 1: Classical cost-aware baselines. Panel B uses the full-trust residual policy ($\kappa = 2$, $\lambda = 1$), with validation-selected $\beta^\star$. In both panels, WW uses a band parameter selected on validation data using the $\kappa = 2$ downside-shaped reward. Entries are paired mean ΔPnL [95% CI] in bp.

A. Classical policies relative to Black–Scholes		
Regime	HW–BS	WW–BS
2018	−1.18 [−3.84, 0.76]	+0.90 [−1.44, 2.91]
2020	+0.44 [−4.60, 4.08]	+0.81 [−4.57, 5.46]
2022	−1.91 [−3.37, −0.43]	−5.09 [−11.96, 1.58]
2024	+1.40 [0.78, 2.09]	+3.67 [1.75, 5.77]
B. Residual policy relative to classical baselines		
Regime	Residual–HW	Residual–WW
2018	+2.00 [−0.10, 4.80]	−0.08 [−2.23, 2.45]
2020	+2.50 [−0.97, 7.53]	+2.13 [−2.93, 7.70]
2022	+3.95 [2.46, 5.46]	+7.14 [0.42, 14.13]
2024	−0.04 [−0.85, 0.75]	−2.32 [−4.35, −0.48]

6.2 Q2: is it regime shift, or sample size?

Takeaway. The within-year oracle has a higher mean than the size-matched historical policy across all three κ values only in 2020. The full-history policy has a lower mean but generally a less adverse lower tail.

Table 2 reports a three-arm diagnostic. The oracle O is trained on alternating start-date batches within the test year and evaluated on the complementary batches. The historical arm S uses an equally sized training sample drawn only from pre-test data. The full-history arm B is the primary expanding-window policy. All three are evaluated on the same held-out batches, with same-date contracts kept together. The oracle is not deployable.

The O and S arms retain B's architecture, objective, optimizer, and validation-selected $\beta^\star$, and are retrained with five random seeds. The contrasts $O - S$ and $B - S$ are diagnostic rather than causal: S only approximates a sample-size control, and the shared $\beta^\star$ may favor B. Within-year training does not guarantee closer distributional matching.

The $O - S$ mean is positive across all three κ values only in 2020, ranging from +4.33 to +5.39 bp, with pointwise 95% intervals excluding zero for $\kappa = 2$ and 3. It is negative throughout 2022 and 2024 and predominantly negative in 2018. The $B - S$ mean is negative in all twelve cells, while B has the least adverse descriptive $\text{CVaR}_{5\%}$ in ten. These results motivate trust control from realized benchmark-relative outcomes.

Table 2: Three-arm diagnostic on common held-out batches. Mean contrasts are in bp with pointwise 95% CIs. Tail reports descriptive benchmark-relative $\text{CVaR}_{5\%}$ in bp.

Regime	κ	$O - S$ [95% CI]	$B - S$ [95% CI]	Tail $O/S/B$
2018	1	−17.12 [−24.56, −9.58]	−20.65 [−29.13, −12.37]	−3.6/−97.9/−1.4
	2	−7.99 [−11.46, −4.56]	−10.34 [−14.52, −6.17]	−1.8/−44.1/−1.5
	3	+0.54 [−0.29, +1.41]	−2.09 [−2.89, −1.27]	−1.8/−10.0/−1.1
2020	1	+4.33 [−0.25, +10.02]	−8.12 [−9.71, −6.28]	−18.9/−13.2/−2.1
	2	+5.39 [+2.53, +8.64]	−4.64 [−5.87, −3.09]	−18.1/−11.6/−1.5
	3	+4.68 [+2.79, +6.57]	−2.23 [−2.86, −1.41]	−24.1/−6.9/−1.0
2022	1	−2.88 [−4.85, −0.87]	−0.44 [−1.48, +0.65]	−1.9/−12.4/−2.6
	2	−1.40 [−2.22, −0.64]	−0.85 [−1.42, −0.29]	−1.0/−5.5/−1.3
	3	−1.09 [−1.68, −0.52]	−0.33 [−0.75, +0.10]	−0.8/−4.1/−0.8
2024	1	−23.10 [−27.13, −18.89]	−20.49 [−23.70, −17.14]	−2.6/−4.1/−0.6
	2	−9.64 [−11.59, −7.67]	−8.69 [−10.17, −7.15]	−1.3/−1.5/−0.3
	3	−5.06 [−6.12, −4.02]	−4.68 [−5.54, −3.81]	−0.6/−1.1/−0.5

6.3 Q3 and Q4: what the guardrail costs, and what it buys

Takeaway. The guardrail is nearly costless when the underlying policy is already stable. For the aggressive policy at $\kappa = 2$, the complete rule gives a realized breach estimate at or below 0.10 in all four regimes, but it is the highest-mean feasible rule only in the 2022 bear regime.

When attenuation is largely unnecessary, its performance cost is small. Under the common $c = 5$ bp budget in 2022 at $\kappa = 1$, the validation-selected policy achieves +5.71 bp under the adaptive rule, compared with +5.75 bp at full trust, with mean trust 0.942 and realized breach frequency 0.0065.

Using the frozen counterfactual ledger without retraining, Table 3 compares our rule with the five deployment baselines across all four held-out regimes. Our rule has a realized breach estimate at or below 0.10 and a positive mean ΔPnL in every regime. In 2022, it achieves the highest mean among rules with observed breach frequency at or below 0.10. Its mean is +22.45 bp, compared with +19.49 bp for CRC, with realized breach frequencies of 0.096 and 0.067, respectively. This comparison uses the same nominal target

Table 3: Published-method-inspired deployment comparison: Full [6], IV [8, 11], ACI [9], CRC [3, 20], and W-CRC [4], against Ours. Entries are mean ΔPnL (bp) / breach for $\beta = 0$, $\kappa = 2$. †: breach > 0.10. Bold indicates the highest mean among rules with observed breach ≤ 0.10.

Regime	Full	IV	ACI	CRC	W-CRC	Ours
2018	+49.69/.166†	**+43.18/.082**	+38.25/.059	+40.97/.090	+32.34/.060	+38.43/.100
2020	+57.41/.149†	+48.50/.110†	**+38.71/.067**	+23.07/.056	+24.36/.042	+35.07/.085
2022	+44.48/.273†	+39.66/.198†	+30.12/.114†	+19.49/.067	+9.60/.034	**+22.45/.096**
2024	**+59.48/.099**	+50.16/.048	+31.52/.020	+53.59/.077	+42.64/.053	+52.93/.082

and monetary budget, while allowing the realized breach frequencies to differ. The highest-mean qualifying rules in 2018, 2020, and 2024 are IV, ACI, and full trust, respectively.

Table 4 complements the fixed-κ rule comparison by applying the complete rule to all twelve regime–κ cells, using the policy-specific validation budgets defined in Section 5.

Table 4: Aggressive $\beta = 0$ policy under the adaptive quantile rule and its policy-specific validation budget. Budget entries give five-seed mean [range]. Mean and CVaR are in bp.

Regime	κ	c mean [range]	Mean	Breach	$\text{CVaR}_{5\%}$
2018	1	34.1[29.7, 36.8]	+40.94	0.099	−128.0
	2	21.8[20.0, 25.2]	+38.43	0.100	−90.8
	3	15.2[11.2, 18.8]	+32.22	0.102	−75.7
2020	1	19.5[18.7, 20.2]	+53.70	0.091	−178.7
	2	9.9[9.1, 11.1]	+35.07	0.085	−92.7
	3	7.6[7.1, 8.1]	+29.20	0.089	−68.8
2022	1	16.2[14.8, 17.7]	+22.81	0.098	−174.5
	2	9.7[8.6, 11.4]	+22.45	0.096	−73.5
	3	8.3[7.7, 8.9]	+21.87	0.097	−54.1
2024	1	27.5[24.5, 32.8]	+58.56	0.078	−82.3
	2	14.9[13.0, 16.5]	+52.93	0.082	−46.9
	3	11.2[9.8, 11.8]	+46.96	0.084	−35.7

Episode-weighted breach-frequency point estimates range from 0.078 to 0.102, with eleven of twelve at or below 0.10. Mean improvements remain positive throughout, ranging from +21.87 to +58.56 bp. Worst-seed episode-weighted frequencies range from 0.081 to 0.105. Date-averaged breach rates range from 0.077 to 0.106. Their one-sided 95% bootstrap upper bounds range from 0.108 to 0.129 and exceed the nominal target of 0.10 in every cell. These results do not establish population-level compliance with the target. The near-target point estimates are consistent with the intended action of the adaptive feedback, but do not isolate the contribution of relevance weighting.

Because the budget is selected separately for each policy–seed pair, this is a within-policy attenuation experiment rather than a like-for-like comparison of training objectives. Section 6.4 therefore holds the deployment budget fixed. The deployment-rule comparison does not include a validation-calibrated constant trust factor and does not compare rules at matched realized breach frequencies. Both comparisons are left for future work.

6.4 Q5: aggressive training at a fixed budget

Takeaway. Under the common 5 bp deployment budget, aggressive training followed by attenuation has a higher held-out mean in all twelve cells. It operates near the 10% breach target, with realized frequencies of 8.6–9.9%, compared with approximately 0–0.7% for the validation-selected policy.

We replay the validation-selected policy C and aggressive policy A under the same specification, $c = 5$ bp and $\varepsilon = 0.10$, with all other deployment choices fixed. Because the evidence ledger stores the counterfactual $\Delta\text{PnL}_i(\lambda)$ at every candidate trust factor, the comparison requires neither retraining nor additional market data.

Table 5 shows that A has the higher sample mean in all twelve regime–κ cells. The paired difference $A - C$ ranges from +2.35 to +26.12 bp, with eleven pointwise 95% intervals excluding zero. The only inconclusive cell is 2022 at $\kappa = 1$, where the estimated difference is +2.35 bp with an interval of $[-5.37, 12.75]$.

Both policies face the same monetary threshold and nominal breach target, but operate at different realized breach frequencies. Table 5 characterizes this trade-off through mean performance, breach frequency, and relative CVaR. The higher sample means establish a common-budget mean advantage, not an advantage at matched breach frequency or a uniform risk improvement.

Table 5: Common-budget comparison at $c = 5$ bp and $\varepsilon = 0.10$. C/A denote the validation-selected and aggressive policies. Intervals use the 10-start-date-batch bootstrap. Breach entries are descriptive frequencies.

Regime	κ	C	A	$A - C$ [95% CI]	Breach (C/A)	$\text{CVaR}_{5\%}$ (C/A)
2018	1	3.07	12.71	+9.64[2.37, 17.33]	.001/.086	−1.7/−59.5
	2	0.78	12.52	+11.74[5.29, 18.39]	.004/.089	−1.3/−45.0
	3	0.65	15.14	+14.49[6.70, 21.56]	.001/.090	−0.9/−42.5
2020	1	8.86	30.06	+21.21[5.81, 36.36]	.005/.095	−2.0/−117.1
	2	2.97	23.71	+20.74[6.25, 34.72]	.001/.096	−0.7/−65.8
	3	1.69	24.07	+22.38[8.21, 35.61]	.000/.095	−0.6/−54.3
2022	1	5.71	8.06	+2.35[−5.37, 12.75]	.007/.092	−2.8/−103.4
	2	2.04	11.99	+9.95[1.03, 22.52]	.001/.096	−1.4/−58.3
	3	1.35	13.88	+12.53[3.73, 22.69]	.000/.097	−0.8/−44.7
2024	1	3.28	20.99	+17.71[9.97, 26.56]	.001/.095	−0.7/−42.4
	2	1.31	27.42	+26.10[16.94, 35.71]	.000/.096	−0.3/−32.0
	3	0.67	26.79	+26.12[17.20, 35.91]	.000/.099	−0.7/−21.6

6.5 Frequency is not severity

Takeaway. The budget c is an exceedance threshold, not a cap on benchmark-relative loss.

In Table 4, the 2020 cell at $\kappa = 1$ has a mean budget of 19.5 bp and a breach frequency of 0.091, but its relative $\text{CVaR}_{5\%}$ is −178.7 bp, approximately nine times the mean budget in magnitude. Under the common 5 bp budget in Table 5, the aggressive policy A has relative $\text{CVaR}_{5\%}$ from −117.1 to −21.6 bp, compared with −2.8 to −0.3 bp for the validation-selected policy C.

The quantile rule scores monetary losses $L_j(\lambda)$. However, the quantile-at-most-c admission test and binary breach feedback impose no separate constraint on exceedance severity. The common-budget mean advantage therefore comes with a substantially heavier lower tail, motivating severity-aware extensions.

6.6 Robustness

Recomputing the lower tail at $\alpha \in \{1\%, 5\%, 10\%\}$ changes its magnitude but not its sign. Moving-block intervals using 5, 10, and 20 start-date batches give the same significance conclusions. At transaction costs of 5, 10, and 20 bp, all 2022 cells remain significant, with

estimated break-even costs of 141–161 bp. Finally, the residual policy raises turnover by only 1.4%–6.8% relative to Black–Scholes, indicating that its advantage is not explained by substantially greater trading intensity.

For the frozen aggressive policy ($\beta = 0, \kappa = 2$), we vary one controller parameter at a time: recency half-life over $\{15, 30, 60\}$ days, IV bandwidth h_{IV} over $\{0.02, 0.04, 0.08\}$, and ACI step size γ over $\{0.01, 0.02, 0.04\}$. The default half-life of 30 days, bandwidth $h_{\mathrm{IV}} = 0.04$, and step size $\gamma = 0.02$ were fixed a priori. The seven configurations, including the shared default, are evaluated across four regimes and five seeds. All 28 configuration–regime cells retain positive mean ΔPnL, with 95% moving-block-bootstrap intervals excluding zero. Breach-frequency sensitivity is not reported, and this analysis does not provide a parameter-tuning rule.

7 Limitations and Conclusion

This paper presents a two-layer conformal trust controller that adapts exposure to a frozen learned policy under regime shift. The main limitations are the frequency-only objective, the lack of finite-sample validity for the weighted adaptive rule, and evaluation limited to short-dated SPXW calls. Across all twelve regime–κ cells, episode-weighted breach-frequency point estimates range from 0.078 to 0.102. For date-averaged breach rates, the one-sided 95% bootstrap upper bounds range from 0.108 to 0.129 and exceed the 0.10 target in every cell. Mean improvements remain positive, but the rule does not uniformly dominate the alternative deployment rules. Under a common budget, aggressive training raises the held-out mean at the cost of a substantially heavier lower tail. Near-target breach frequencies can coexist with severe benchmark-relative losses: the budget is an exceedance threshold, not a loss cap.

The broader interest of the construction lies in this separation: Layer 1 is trained once and frozen, while residual exposure is selected afterwards from completed outcomes. For the same benchmark-relative loss definition, the monetary breach threshold c and deployment guardrail can be revised without retraining the policy. Where the four requirements in Section 3.3 hold, this suggests an alternative to detect–then–retrain: completed evidence determines how much of a frozen policy to run. Future work includes severity-aware deployment control and evaluation beyond hedging.

Acknowledgments

This work is supported by the United Kingdom EPSRC (grant numbers UKRI256, EP/V028251/1, EP/N031768/1, EP/S030069/1, and EP/X036006/1), AMD, Altera, and Intel.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. PMLR, 22–31. https://proceedings.mlr.press/v70/achiam17a.html

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe Reinforcement Learning via Shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. AAAI Press, 2669–2678. doi:10.1609/aaai.v32i1.11797

[3] Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal Risk Control. In *International Conference on Learning Representations*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (Eds.), Vol. 2024. 55198–55218. https://proceedings.iclr.cc/paper_files/paper/2024/file/f3549ef9b5ff520a7e41ff3cc306ab2b-Paper-Conference.pdf

[4] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics* 51, 2 (April 2023), 816–845. doi:10.1214/23-AOS2276

[5] Fischer Black and Myron Scholes. 1973. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81, 3 (1973), 637–654. http://www.jstor.org/stable/1831029

[6] H. Buehler, L. Gonon, J. Teichmann, and B. Wood. 2019. Deep hedging. *Quantitative Finance* 19, 8 (2019), 1271–1291. doi:10.1080/14697688.2019.1571683

[7] Jay Cao, Jacky Chen, Soroush Farghadani, John Hull, Zissis Poulos, Zeyu Wang, and Jun Yuan. 2023. Gamma and vega hedging using deep distributional reinforcement learning. *Frontiers in Artificial Intelligence* 6, Article 1129370 (Feb. 2023), 11 pages. doi:10.3389/frai.2023.1129370

[8] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.* 46, 4, Article 44 (March 2014), 37 pages. doi:10.1145/2523813

[9] Isaac Gibbs and Emmanuel Candes. 2021. Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 1660–1672. https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf

[10] Edwin Hamel-De le Court, Francesco Belardinelli, and Alexander W. Goodall. 2025. Probabilistic Shielding for Safe Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI Press, 16091–16099. doi:10.1609/aaai.v39i15.33767

[11] James D. Hamilton. 1989. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* 57, 2 (1989), 357–384. http://www.jstor.org/stable/1912559

[12] Steven L. Heston. 1993. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies* 6, 2 (04 1993), 327–343. doi:10.1093/rfs/6.2.327

[13] John Hull and Alan White. 2017. Optimal delta hedging for options. *Journal of Banking & Finance* 82 (2017), 180–190. doi:10.1016/j.jbankfin.2017.05.006

[14] Jordan Lekeufack, Anastasios N. Angelopoulos, Andrea Bajcsy, Michael I. Jordan, and Jitendra Malik. 2024. Conformal Decision Theory: Safe Autonomous Decisions from Imperfect Predictions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. doi:10.1109/ICRA57147.2024.10610041

[15] Oskari Mikkilä and Juho Kanniainen. 2023. Empirical deep hedging. *Quantitative Finance* 23, 1 (2023), 111–122. doi:10.1080/14697688.2022.2136037

[16] Kristina Miller, Christopher K. Zeitler, William Shen, Kerianne Hobbs, John Schierman, Mahesh Viswanathan, and Sayan Mitra. 2024. Optimal Runtime Assurance via Reinforcement Learning. In *Proceedings of the 15th ACM/IEEE International Conference on Cyber-Physical Systems*. IEEE, 67–76. doi:10.1109/ICCPS61052.2024.00013

[17] Dung T. Phan, Radu Grosu, Nils Jansen, Nicola Paoletti, Scott A. Smolka, and Scott D. Stoller. 2020. Neural Simplex Architecture. In *NASA Formal Methods (Lecture Notes in Computer Science, Vol. 12229)*. Springer, 97–114. doi:10.1007/978-3-030-55754-6_6

[18] Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Don Stanton. 2026. Conformal Policy Control. In *Proceedings of the 43rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 306)*. PMLR, 99772–99824. https://proceedings.mlr.press/v306/prinster26a.html

[19] R.Tyrrell Rockafellar and Stanislav Uryasev. 2002. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance* 26, 7 (2002), 1443–1471. doi:10.1016/S0378-4266(02)00271-6

[20] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2022. *Algorithmic Learning in a Random World*. Springer, Cham. doi:10.1007/978-3-031-06649-8

[21] A. E. Whalley and P. Wilmott. 1997. An Asymptotic Analysis of an Optimal Hedging Model for Option Pricing with Transaction Costs. *Mathematical Finance* 7, 3 (1997), 307–324. doi:10.1111/1467-9965.00034