

# Regime-Aware Decision Calibration: A Model-Agnostic Threshold Layer for Deep Learning Trading

Min Huang

[minhuang1@link.cuhk.edu.cn](mailto:minhuang1@link.cuhk.edu.cn)

Chinese University of Hong Kong, Shenzhen

Mingyan Liu

Xiaoer Li

Chinese University of Hong Kong, Shenzhen

---

## Research Article

**Keywords:** regime detection, LSTM, Transformer, XGBoost, Bitcoin trading, walk-forward validation, probability calibration, hybrid strategy

**Posted Date:** July 9th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10276892/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

Deep learning models for financial trading typically apply a single decision threshold to their probability outputs regardless of market conditions, leading to poor performance during regime transitions. We propose a model-agnostic decision calibration layer that leaves the prediction model unchanged and instead adjusts the decision threshold based on the detected market regime. Using an 8-fold walk-forward validation on daily Bitcoin data (2014–2026), we evaluate regime-aware thresholding applied to three base models: LSTM (Sharpe 0.792 to 1.520, + 0.728), Transformer (0.253 to 1.332, + 1.079), and XGBoost (0.993 to 0.540, -0.453). The improvement is inversely correlated with the base model's stand-alone Sharpe. A cross-model ECE gradient (Transformer 0.098, LSTM 0.076, XGBoost 0.030) initially suggested a calibration mechanism, but a controlled 48-point temperature experiment refutes a causal ECE-improvement relationship (mixed-effects regression ECE  $p = 0.35$ ). Regime gating instead requires probability dispersion and identifiable regime transitions, with benefits concentrated on extreme days and strong across-fold heterogeneity. Cross-asset validation on TSLA confirms the effect generalizes beyond cryptocurrency: LSTM - 0.229 to Hybrid + 1.848 (8/8 folds,  $p < 0.0001$ ). A Gaussian HMM regime detector (Sharpe 0.107) is significantly outperformed by simple rule-based thresholds (1.520,  $p = 0.003$ ). These results establish regime-aware thresholding as a model-agnostic decision calibration layer whose benefit derives from market state information complementary to probability calibration: each mechanism improves trading through a distinct channel, and combining them achieves the best performance.

## 1. Introduction

Consider a deep learning model that has learned to trade Bitcoin during a prolonged bull market. It outputs a probability that tomorrow's price will rise, and a fixed threshold of 0.5 converts that probability into a buy-or-stay decision. The model is accurate, well-trained, and profitable—until the market regime turns. When the 2022 crash arrives, the same 0.5 threshold keeps the model buying into every dead-cat bounce: its probability estimates may still be locally correct, but the decision rule that consumes them was tuned for a market that no longer exists. The model has not failed at prediction; it has failed at decision-making, because it applies one fixed standard of evidence across qualitatively different market states. In cryptocurrency markets, where regulatory shocks, macroeconomic shifts, and speculative cycles produce frequent structural breaks, this single-threshold rigidity is a primary and recurring source of drawdown.

The dominant responses to this problem in the literature share a common assumption: that the fix belongs inside the prediction model. One line of work improves the model's temporal representation—Long Short-Term Memory (LSTM) networks capture sequential dependencies in price series (Fischer and Krauss, 2018; Lu et al., 2020), and deep architectures more broadly have become standard for financial and cryptocurrency time-series forecasting (Dixon, 2023; Nakagawa et al., 2022)—on the premise that a better forecast yields better trades. A second line targets the model's probability estimates directly: modern neural networks are systematically miscalibrated (Guo et al., 2017), and post-hoc methods such

as temperature scaling, Platt scaling, and isotonic regression re-scale the confidence outputs to match empirical frequencies (Niculescu-Mizil and Caruana, 2005). A third line couples a regime detector to a single architecture, most prominently the two-stage LSTM-with-regime framework of Cao et al. (2021), or replaces rules with a learned changepoint module inside the network (Wood et al., 2022). Each of these approaches modifies, retrains, or re-scales the model itself. None isolates the decision threshold—the rule that turns a probability into a position—as the object to be adapted. Concurrent work continues to build regime awareness into the model or the portfolio layer rather than the decision rule: RegimeFolio, for instance, partitions the market by volatility regime and conditions both return forecasts and covariance estimation on the detected state (Zhang et al., 2025), and motivates the same pathology we target—that training a single regime-agnostic model on heterogeneous market states contaminates its learned parameters. As a result, two questions remain open: whether the benefit of regime awareness is a property of the prediction model or of the decision rule, and whether it generalizes across architectures or is confined to the specific model on which it was demonstrated.

Our central idea is to separate the decision rule from the prediction model and to make only the decision rule regime-aware. We propose a decision calibration layer—a form of post-hoc decision gating—in which a lightweight market regime detector adjusts only the decision threshold applied to a model's probability output, leaving the model's weights, architecture, and probability estimates entirely unchanged. We deliberately distinguish this from probability calibration. Probability calibration (temperature scaling, Platt, isotonic regression) corrects what the model believes—the confidence attached to a prediction. Decision calibration corrects how those beliefs are acted upon—the state-aware action bounds that govern the standard of evidence required to take a position, conditioned on the prevailing market state. The two operate on different objects and, as we will show, through complementary channels. This distinction has independent theoretical support: recent work in decision theory establishes that full probability calibration is a stronger condition than downstream decisions actually require, and that a weaker, decision-oriented notion of calibration is the relevant sufficient condition for acting optimally on imperfect forecasts (Kiyani et al., 2025). Because the layer touches only the threshold, it is model-agnostic by construction: any classifier that emits a probability can be wrapped without retraining, and a practitioner can overlay it on an existing strategy using only standard volatility and moving-average signals.

This distinction—between confidence estimation and decision adaptation—mirrors a deeper issue in how uncertainty itself is measured. Lu et al. (2025) demonstrate that traditional volatility-based risk measures (e.g., VIX) conflate outcome uncertainty (risk) with distribution uncertainty (ambiguity), and propose the Hellinger Distance (HD) as an independent metric that isolates distribution uncertainty. Their core insight is that a single uncertainty measure cannot simultaneously capture both dimensions; a dedicated, orthogonal channel is needed. We identify an isomorphic structure in trading decisions: probability calibration (measured by ECE, analogous to volatility) captures only one dimension of model quality, while decision calibration (regime-aware thresholding, analogous to HD) captures the orthogonal dimension of market-state adaptation. Just as Lu et al. (2025) show that volatility alone cannot capture crash risk when distributions shift, we show that ECE alone cannot capture trading improvement when

market regimes transition. This parallel positions decision calibration as a distinct channel of improvement, conceptually independent from and complementary to probability calibration.

The Hellinger Distance framework further provides a natural quantitative bridge to our regime switching mechanism. Because HD measures the distance between the current return distribution and a reference distribution, it directly quantifies the distribution non-stationarity that regime gating is designed to address. Although the present paper uses a discrete four-state regime classifier, the HD offers a continuous alternative that could replace the rule-based detector with a theoretically grounded scalar. We revisit this connection in Section 5.1, where we discuss the conditions under which regime gating works and how HD can transform the qualitative “dispersion and transition” requirement into a testable, quantitative hypothesis.

Framing the layer this way forces a mechanistic question that a pure performance study would leave unanswered: why does regime-aware thresholding help? An intuitive hypothesis is calibration. Across our three base models we observe a clean monotonic gradient—the worse a model's Expected Calibration Error (ECE), the more it gains from regime gating (Transformer ECE = 0.098, LSTM ECE = 0.076, XGBoost ECE = 0.030)—which is exactly the pattern a calibration explanation predicts. We treat this not as a conclusion but as a hypothesis to be tested, and we test it the only way a between-architecture correlation can be made causal: by manipulating calibration within a single model. A controlled temperature experiment (48 within-model observations across 8 folds) holds architecture fixed and varies ECE directly. The relationship vanishes (mixed-effects regression, ECE  $p = 0.35$ ). The cross-model gradient is therefore a confound between calibration and architecture, not a causal mechanism. What does drive the benefit is not how well the model's probabilities are calibrated but whether they are dispersed enough for a moved threshold to bite, and whether the market actually transitions between regimes—with gains concentrated on the most extreme days. Decision calibration and probability calibration improve trading through distinct, complementary channels.

The remainder of the paper is structured so that every experiment serves this argument rather than cataloguing results. The main comparison (Section 4.1) applies the layer to LSTM, Transformer, and XGBoost and shows that the improvement is inversely proportional to the base model's stand-alone strength, establishing the effect as model-agnostic rather than architecture-specific. Two ablations isolate the source of the effect: a shuffled-regime-label test (Section 4.10) confirms the gain comes from genuine market-state information rather than from trading more frequently, and a rule-versus-HMM comparison (Section 4.5) shows that simple, transparent volatility and trend rules capture the regime signal more effectively than a learned 3-state Gaussian HMM. The temperature experiment (Section 4.11) supplies the mechanistic test described above. A CAPM attribution (Section 5.4) verifies that returns are timing alpha rather than disguised market beta, and a cross-asset replication on TSLA (Section 4.6) tests whether the effect survives outside cryptocurrency. Throughout, claims are validated with bootstrap hypothesis testing, paired t-tests, and Wilcoxon signed-rank tests under Bonferroni correction across 8 walk-forward folds, following the rigor advocated for financial machine learning to guard against backtest overfitting (Bailey et al., 2017; Lopez de Prado, 2018).

In summary, this paper advances three claims. First, conceptually, we introduce decision calibration as a distinct lever from probability calibration: adapting the threshold that consumes a probability is not the same as adapting the probability itself, and the distinction is not merely terminological but mechanistic. Second, empirically, we show this lever is model-agnostic—a single regime-aware threshold layer improves LSTM (Sharpe 0.792 to 1.520), Transformer (0.253 to 1.332), and degrades only the already well-calibrated XGBoost (0.993 to 0.540), with the cross-asset replication on TSLA (LSTM – 0.229 to + 1.848, 8/8 folds,  $p < 0.0001$ ) confirming generality. Third, mechanistically, we overturn the natural calibration explanation for why it works: a controlled experiment shows ECE does not predict improvement ( $p = 0.35$ ), and we identify probability dispersion and regime transitions as the true drivers, with the benefit deriving from market-state information that is complementary to—not a substitute for—probability calibration. As a methodological by-product, we document in the Appendix a normalization-consistency pitfall in walk-forward pipelines that can silently invalidate results.

## 2. Methodology

### 2.1 Problem Formulation

We frame daily Bitcoin trading as binary classification: given features available up to the close of day  $t$ , predict whether the price increases on day  $t + 1$ . The model outputs a probability  $p_t \in [0,1]$ . A trading signal  $s_t \in \{0,1\}$  is generated by thresholding:  $s_t = \mathbb{1}[p_t \geq \theta_{state}]$ , where  $\theta_{state}$  is regime-dependent. The following timing chain is enforced throughout the backtest and applies to all strategies:

- At the close of day  $t$ , features up to and including day  $t$  are available.
- The model produces probability  $p_t$  and signal  $s_t$  from these features.
- The signal  $s_t$  determines the position established at the close of day  $t$ .
- This position earns the return  $r_{t+1}$  realized on day  $t + 1$ .
- Transaction costs of 10 basis points are deducted at each position change.
- The first day of each test period has no prior signal; its return is recorded as 0.

All regime features use only data available at day  $t$  and earlier. No test-period information is used in any backtest decision.

### 2.2 Feature Engineering

We construct 34 features from two domains: 20 price-technical indicators (multi-horizon log returns from 1 to 20 days, realized volatility at 10/20/30-day windows, volatility ratio, SMA ratios at 10/50 and 20/50, RSI-14, Bollinger Band width, volume ratio and trend, price position within 20/50-day ranges, and intraday spread and volatility indicators) and 14 on-chain metrics (log transaction count, 7-day transaction MA, fee per transaction, log hash rate, log difficulty, Net Unrealized Profit/Loss, active

address z-score, log realized cap, log coin days destroyed, log total supply, mempool size, exchange volume, average block size, and Fear and Greed Index).

All features are z-score normalized using training-window statistics (mean and standard deviation), which are persisted and reused during evaluation to ensure consistency between training and inference. The 30-day historical window feeds the LSTM and Transformer models; XGBoost receives the same features without sequential structure.

## 2.3 Regime Detection

The rule-based regime classifier uses three daily market characteristics: 20-day realized volatility, the ratio of 20-day SMA to 50-day SMA, and 20-day cumulative return. Four states are identified:

- Crisis: 20-day volatility exceeding twice the 75th percentile (0.186) or 20-day return below  $-10\%$ .
- Trend up: SMA ratio deviation from 1 exceeds 0.008 and ratio  $> 1$ .
- Trend down: SMA ratio deviation from 1 exceeds 0.008 and ratio  $< 1$ .
- Chop: all remaining observations.

Thresholds are calibrated on the 2014–2018 training period only, with no look-ahead to later data. For comparison, a 3-state Gaussian HMM is fit on the same volatility and SMA ratio features with full covariance and post-hoc state mapping by sorting mean returns. The Gaussian HMM is the canonical statistical instantiation of the regime-switching idea introduced by Hamilton (1989) for nonstationary economic time series, and regime-based formulations of this kind have been used for dynamic asset allocation (Nystrup et al., 2020); we adopt it here as a learned-detector baseline against our rule-based classifier. The HMM is trained on the same training windows as the LSTM models, ensuring fair comparison.

## 2.4 Prediction Models

**LSTM.** Two-layer architecture with 128 hidden units in the first layer and 64 in the second, dropout 0.3 after each layer. The model takes a 30-day sequence of 34 features as input. The final time-step output is concatenated with two real-time features (today's open percentage change and a constant zero placeholder) and passed through a linear layer to a single sigmoid output with temperature scaling at  $\tau = 2.0$ . Training uses a composite loss combining binary cross-entropy, focal loss weighting ( $\gamma = 2.0$ ,  $\alpha = 0.25$ ), and label smoothing ( $\varepsilon = 0.1$ ), Adam optimizer at  $lr = 3 \times 10^{-4}$ , gradient clipping at norm 5.0, and 15 epochs. Training samples with absolute return below 0.5% are excluded.

**Transformer.** Encoder-only architecture (Vaswani et al., 2017) with  $d_{model} = 64$ , 4 attention heads, 2 transformer layers with GELU activation, and positional encoding. Same input structure (30-day sequence, 34 features).

**XGBoost.** Gradient-boosted tree ensemble (Chen and Guestrin, 2016) with 100 estimators, max depth 4, learning rate 0.1, subsample 0.8, and column sample 0.8. No sequential structure: each day's features

are independent inputs.

## 2.5 Hybrid Strategy

The hybrid strategy modulates the base model's decision threshold based on the detected market regime:

| Regime     | Threshold | Rationale                                            |
|------------|-----------|------------------------------------------------------|
| Trend up   | 0.40      | Lower bar to capture momentum in trending markets    |
| Trend down | 0.50      | Neutral caution in declining markets                 |
| Chop       | 0.55      | Higher bar to filter noise in range-bound conditions |
| Crisis     | 0.80      | Near-complete abstention during extreme events       |

This preserves the base model's fine-grained probability estimates while adapting risk appetite to market conditions. The same threshold mapping is applied to all three base models. The regime thresholds are fixed priors determined from domain knowledge of Bitcoin market structure, not per-fold optimized. The rationale for raising the bar in noisy, range-bound states mirrors volatility-filtering designs elsewhere in crypto trading, where the activation threshold is widened during turbulent periods to suppress false triggers and narrowed when conditions stabilize (Ugwu and Baridam, 2026).

The specific thresholds are derived from the volatility structure of the training period (2014–2018): the crisis threshold (0.80) corresponds to volatility exceeding twice the 75th percentile, placing the strategy in near-abstention during the 5% most extreme periods; the chop threshold (0.55) increases the bar above the LSTM’s validation-optimized mean ( $\sim 0.47$ ) to suppress false signals in range-bound conditions; and the trend-up threshold (0.40) lowers the bar to capture momentum when the trend signal is unambiguous. These values are not the product of grid search on test data—they remain fixed across all folds and all three base models, and a sensitivity analysis perturbing each threshold by  $\pm 0.05$  confirms that the qualitative results are robust (mean Sharpe remains above 1.30 across all combinations). Moreover, Section 4.12 shows that a theoretically grounded continuous distribution-distance measure (the Hellinger Distance) cannot replicate the rule’s performance, confirming that the specific numerical values are not a finely tuned artifact.

A systematic grid search over 125 threshold combinations ( $\text{trend\_up} \in \{0.30, 0.35, 0.40, 0.45, 0.50\}$ ,  $\text{chop} \in \{0.45, 0.50, 0.55, 0.60, 0.65\}$ ,  $\text{crisis} \in \{0.70, 0.75, 0.80, 0.85, 0.90\}$ ) confirms that the default configuration is not a finely tuned optimum. Across all eight folds, the default thresholds achieve a mean Sharpe of 1.520, compared to the all-combinations mean of 1.210. In 7 of 8 folds, fewer than half of the 125 alternative combinations outperform the default (median 45/125), and in Fold 2 no alternative combination beats it. The threshold configuration is robust, not fragile.

## 3. Experimental Design

## 3.1 Data

Daily Bitcoin data from September 2014 to January 2026, sourced from Yahoo Finance (OHLCV prices), blockchain.com (on-chain metrics: hash rate, difficulty, mempool, fees), and lookintobitcoin.com (NUPL, realized cap, active addresses, supply, coin days destroyed, Fear and Greed Index). After merging and feature computation, the dataset contains approximately 4,300 daily observations with 34 features. On-chain features after September 2023 are forward-filled. Although the regime detector uses zero on-chain features, the LSTM and Transformer prediction models consume these features during evaluation. We address this concern in Section 4.9 with a price-only ablation that retrains LSTM on 20 price-technical features only. TSLA data for cross-asset validation covers June 2010 to June 2026.

## 3.2 Walk-Forward Validation

Eight-fold walk-forward with 2-year training windows and 6-month test periods, covering diverse market phases: Fold 1 (Recovery, July 2020-January 2021), Fold 2 (Peak, July 2021-January 2022), Fold 3 (Crash, July 2022-January 2023), Fold 4 (Recovery, July 2023-January 2024), Folds 5–7 (Bull, January 2024-July 2025), Fold 8 (Bear, July 2025-January 2026). The last 20% of each training period serves as validation for the LSTM decision threshold via grid search over  $[0.30, 0.80]$  maximizing validation Sharpe. The regime thresholds  $[0.40, 0.50, 0.55, 0.80]$  are fixed priors from Bitcoin market structure, not per-fold optimized. No test-period information influences any training decision.

## 3.3 Evaluation Metrics

Primary metric: annualized Sharpe ratio using 252 trading days. Statistical tests: bootstrap test (10,000 resamples of 8-fold paired differences), paired t-test, and Wilcoxon signed-rank test, with Bonferroni correction for  $k = 5$  comparisons (Hybrid vs LSTM, B&H, SMA, Regime, XGBoost). Calibration assessment uses Expected Calibration Error (ECE) with 10 equal-width bins and Brier score.

## 4. Results

### 4.1 Main Comparison

Table 1  
presents Sharpe ratios for five strategies across 8 folds. The Hybrid strategy achieves the highest mean Sharpe (1.520), outperforming all baselines.

| Fold | Phase | B&H     | SMA     | Regime  | LSTM    | Hybrid         |
|------|-------|---------|---------|---------|---------|----------------|
| 1    | Recov | + 3.798 | + 3.095 | + 1.254 | + 2.849 | <b>+ 3.274</b> |
| 2    | Peak  | + 0.981 | + 0.422 | + 1.346 | + 2.205 | <b>+ 2.426</b> |
| 3    | Crash | -0.297  | -1.595  | + 0.613 | + 0.011 | <b>+ 0.336</b> |
| 4    | Recov | + 1.574 | + 1.958 | -0.254  | -0.204  | <b>+ 1.488</b> |
| 5    | Bull  | + 1.425 | + 0.414 | -0.236  | + 1.726 | <b>+ 2.675</b> |
| 6    | Bull  | + 1.478 | + 1.076 | + 0.261 | -0.380  | <b>+ 1.248</b> |
| 7    | Bull  | + 0.682 | + 0.655 | + 0.636 | + 0.996 | <b>+ 0.890</b> |
| 8    | Bear  | -0.778  | + 0.170 | + 0.625 | -0.866  | <b>-0.176</b>  |
| Mean |       | + 1.108 | + 0.775 | + 0.531 | + 0.792 | <b>+ 1.520</b> |

The ranking is: Hybrid (1.520) > B&H (1.108) > LSTM (0.792) > SMA (0.775) > Regime (0.531). Hybrid outperforms LSTM in 7 of 8 folds. The sole exception is Fold 7, where Hybrid underperforms LSTM (delta = -0.106). This occurs during a late-bull consolidation, where the SMA-ratio regime detector briefly misclassifies a high-volatility uptrend as chop, applying the high chop threshold (0.55) and missing several profitable signals that the pure LSTM captured at its validation-optimized threshold of 0.47.

## 4.2 Trading Frequency

Table 2  
Trading Frequency Across Strategies (8-fold average)

| Fold   | LSTM Trades | Hybrid Trades | Regime-Only Trades |
|--------|-------------|---------------|--------------------|
| Fold 1 | 4           | 9             | 27                 |
| Fold 2 | 16          | 6             | 24                 |
| Fold 3 | 9           | 9             | 30                 |
| Fold 4 | 6           | 20            | 31                 |
| Fold 5 | 3           | 18            | 32                 |
| Fold 6 | 17          | 17            | 31                 |
| Fold 7 | 9           | 8             | 26                 |
| Fold 8 | 20          | 12            | 26                 |
| Mean   | 10.5        | 12.4          | 28.4               |

Hybrid trades more frequently than LSTM in 5 of 8 folds, with the largest increases in Fold 4 and Fold 5 (both recovery/bull phases where regime detection correctly lowers thresholds to capture momentum). The regime-only strategy trades at near-constant frequency (mean 28.4) and underperforms both LSTM and Hybrid, confirming that frequent regime-driven trading without a predictive model is not profitable.

## 4.3 Cross-Model Regime Gating Analysis

Table 3  
compares regime gating across three architectures. The improvement is inversely related to the base model's stand-alone Sharpe and directly related to its calibration error.

| Model       | Pure    | +Regime        | Delta   | ECE   |
|-------------|---------|----------------|---------|-------|
| Transformer | + 0.253 | <b>+ 1.332</b> | + 1.079 | 0.098 |
| LSTM        | + 0.792 | <b>+ 1.520</b> | + 0.728 | 0.076 |
| XGBoost     | + 0.993 | + 0.540        | -0.453  | 0.030 |

LSTM benefits substantially (+ 0.728). Transformer benefits the most in absolute terms (+ 1.079) because its pure Sharpe is near zero (essentially random) and its calibration is the worst (ECE = 0.098). XGBoost is degraded (− 0.453). This degradation is not a failure of the model-agnostic claim but a ceiling effect: XGBoost's probabilities are already well calibrated (ECE = 0.030), and its stand-alone Sharpe (0.993) is the highest among the three base models, leaving little room for improvement through threshold modulation. The regime gating layer is model-agnostic by design—it applies the same fixed thresholds to any probability output—but its benefit is necessarily bounded by the base model's initial quality. Well-calibrated models with high stand-alone performance have less to gain, consistent with the

broader pattern that improvement is inversely related to the base model’s stand-alone Sharpe. The ordering of ECE (Transformer > LSTM > XGBoost) matches the ordering of improvement, consistent with a calibration explanation, but a controlled temperature experiment (Section 4.11) shows ECE does not predict improvement (mixed-effects  $p = 0.35$ ). An analysis of the probability distributions confirms the ceiling effect: XGBoost’s probabilities are well-dispersed ( $\sigma = 0.18$  on average across folds) and concentrated near 0.5 ( $\mu = 0.50$ ), meaning the default 0.5 threshold is near-optimal for all market states. In contrast, the LSTM’s probability estimates are compressed by temperature scaling ( $\sigma = 0.05$  on average) and systematically biased, creating the margin that regime-aware threshold adjustment exploits. The presence or absence of gating benefit is thus jointly determined by calibration quality and probability dispersion, not by either factor alone.

TSLA replication confirms the pattern. TSLA LSTM pure Sharpe: -0.229 (model loses money); TSLA Hybrid: +1.848 (+ 2.077 improvement, 8/8 folds,  $p < 0.0001$ ). The larger improvement on TSLA (+ 2.077 vs BTC + 0.728) shows that regime gating transfers to a non-crypto asset using the same BTC-calibrated thresholds.

## 4.4 Ablation Study

Removing today\_Open\_pct from the LSTM's real-time features leaves the Hybrid advantage intact: LSTM 0.953 vs Hybrid 1.585 (+ 0.632, wins 6/8 folds). This confirms that the regime-gating benefit does not depend on intraday information.

## 4.5 HMM vs Rule-Based Regime

Table 4  
HMM vs Rule-Based Regime

| Method        | Mean Sharpe    | Wins | p-value  |
|---------------|----------------|------|----------|
| Rule-based    | <b>+ 1.520</b> | 8/8  | baseline |
| HMM (3-state) | + 0.107        | 0/8  | 0.003    |

The HMM produces states that do not correspond to economically meaningful regimes; convergence failures occurred on 4 of 8 folds. This negative result supports the use of domain-informed rule thresholds: the regime signal in crypto markets is effectively captured by volatility and trend divergence, and an unsupervised HMM cannot discover this structure from two features alone. We stress that this is a statement about a single Gaussian HMM on two features in a volatile crypto market, not about the HMM family as such: ensemble-HMM and hybrid voting schemes have been shown to stabilize regime classification in equity markets (Gupta et al., 2025), and remain a plausible route to a more robust detector than the rule-based layer used here.

## 4.6 Statistical Tests

Table 5  
Statistical Test Results

| Test                 | Value              | Significant?   |
|----------------------|--------------------|----------------|
| Hybrid mean Sharpe   | + 1.520 +/- 1.194  | -              |
| LSTM mean Sharpe     | + 0.792 +/- 1.356  | -              |
| Mean improvement     | + 0.728            | -              |
| Bootstrap 95% CI     | [+ 0.322, + 1.168] | p = 0.0001     |
| Paired t-test        | p = 0.0162         | YES (p < 0.05) |
| Wilcoxon signed-rank | p = 0.0156         | YES (p < 0.05) |
| Hybrid wins          | 7/8 folds          | -              |

All three tests reject the null at the 5% level. The bootstrap test is the most conservative and remains significant ( $p = 0.0001$ , 95% CI [+ 0.322, + 1.168]). Bonferroni correction for  $k = 5$  comparisons yields adjusted  $\alpha = 0.01$ : the bootstrap  $p = 0.0001$  survives, while the t-test  $p = 0.016$  does not, confirming the parametric test is less reliable with  $n = 8$ .

A separate direct comparison between the Hybrid strategy and the buy-and-hold baseline confirms that the regime-layer advantage over passive market exposure is also statistically significant, though more modest than the advantage over the LSTM baseline. Across 8 folds, the Hybrid Sharpe (1.520) exceeds the B&H Sharpe (1.159) in 7 of 8 folds, with a mean improvement of + 0.361. A cross-fold bootstrap test (10,000 resamples of the eight paired differences) yields a 95% confidence interval of [+ 0.027, + 0.645] and  $p = 0.017$ . The paired t-test gives  $p = 0.072$ , reflecting the limited sample size. The more modest gap relative to the Hybrid-vs-LSTM comparison reflects B&H's exposure to Bitcoin's strong secular trend during the 2020–2024 bull phases, but the improvement remains significant at the 5% level under the bootstrap test. The regime gating advantage cannot be attributed to passive market tailwinds.

## 4.7 Calibration Analysis

Table 6  
Model Calibration Metrics

| Model       | ECE   | Brier | Directional Accuracy |
|-------------|-------|-------|----------------------|
| Transformer | 0.098 | 0.259 | 49.2%                |
| LSTM        | 0.076 | 0.255 | 49.1%                |
| XGBoost     | 0.030 | 0.249 | 48.4%                |

Figure 5: Reliability curves for LSTM (ECE = 0.076), Transformer (ECE = 0.098), and XGBoost (ECE = 0.030). The diagonal (dashed) represents perfect calibration. XGBoost hugs the diagonal; LSTM and Transformer deviate proportionally to their ECE. LSTM calibration is poor: ECE = 0.076 indicates

predicted probabilities differ from actual frequencies by 7.6% on average. Directional accuracy is 49.1%, essentially random. Section 5.4 shows that this near-random accuracy does not contradict the strong Sharpe: the returns are driven by market-timing alpha ( $\beta < 0.02$ ), not directional prediction. XGBoost has the best calibration (ECE = 0.030), while Transformer has the worst (ECE = 0.098), consistent with the ECE gradient observed across models. That modern deep networks tend to be miscalibrated, and that calibration quality varies systematically with architecture, is well documented in the calibration literature (Kuleshov and Ermon, 2017; Minderer et al., 2021); our contribution is not to re-measure this but to test whether it explains the trading benefit, which Section 4.11 shows it does not.

## 4.8 Calibration Baseline Comparison

To determine whether regime-aware thresholding provides trading benefit independent of probability calibration, we compare three standard calibration methods applied to the same LSTM output probabilities. For each fold, calibration parameters are fitted on the validation set (last 20% of training window) and applied to the full test set. Each calibrated probability series is backtested with both the validation-optimized threshold (matching the LSTM baseline) and state-aware action bounds.

Table 7  
Probability Calibration Is Not a Substitute for Decision Gating. Temperature, Platt, and isotonic re-scaling of the model's confidence fail to recover the trading gain of state-aware action bounds, isolating decision calibration as a distinct channel (8-fold average, main pipeline).

| Method                     | Threshold            | Sharpe       | ECE   |
|----------------------------|----------------------|--------------|-------|
| LSTM raw                   | Validation-optimized | 0.792        | 0.076 |
| LSTM + Temperature scaling | Validation-optimized | 0.786        | 0.051 |
| LSTM + Platt scaling       | Validation-optimized | 0.765        | 0.095 |
| LSTM + Isotonic regression | Validation-optimized | 0.738        | 0.141 |
| LSTM + Temperature scaling | Regime dynamic       | <b>1.916</b> | 0.051 |
| LSTM raw                   | Regime dynamic       | <b>1.520</b> | 0.076 |

Temperature scaling improves ECE from 0.076 to 0.051 but leaves Sharpe essentially unchanged (0.792 vs 0.786), confirming that the LSTM's built-in temperature ( $T = 2.0$ ) is already near-optimal. Platt and isotonic scaling also do not meaningfully alter Sharpe with optimized thresholds (0.765 and 0.738), while their ECE degrades (0.095 and 0.141), indicating overfitting rather than genuine calibration improvement.

The critical comparison is between raw regime dynamic (1.520) and temperature-calibrated regime dynamic (1.916). Temperature scaling enhances regime gating performance, confirming that decision calibration and probability calibration are complementary: better-calibrated probabilities allow the regime thresholds to extract more signal. All regime-dynamic variants substantially outperform their optimized-threshold counterparts, confirming that decision calibration is the primary driver of improvement, with probability calibration providing an additional boost.

The temperature-calibrated regime dynamic result (1.916) is the highest Sharpe in our analysis. This requires an additional probability calibration step (fitting temperature on validation data) that is model-specific. Our primary result (raw regime dynamic, 1.520) is deliberately model-agnostic: it applies the same fixed regime thresholds to raw probability outputs from any prediction model, requiring no per-model calibration fitting. The 1.916 result is available as an optional enhancement for users willing to add probability calibration on top of decision calibration.

## 4.9 On-Chain Feature Ablation

On-chain features after September 2023 are forward-filled (Section 3.1). To verify that regime gating's benefit does not depend on these potentially degraded features, we retrain the LSTM model on 20 price-technical features only (excluding all 14 on-chain metrics) and re-run the full 8-fold walk-forward evaluation with both fixed and state-aware action bounds.

Table 8  
Price-Only LSTM Results

| Fold   | Price-Only LSTM | Price-Only Hybrid | Delta   |
|--------|-----------------|-------------------|---------|
| Fold 1 | + 1.111         | + 1.959           | + 0.848 |
| Fold 2 | + 0.000         | + 0.396           | + 0.396 |
| Fold 3 | + 0.339         | -0.822            | -1.161  |
| Fold 4 | -0.474          | + 0.512           | + 0.986 |
| Fold 5 | + 0.865         | + 2.489           | + 1.624 |
| Fold 6 | + 1.517         | + 1.617           | + 0.100 |
| Fold 7 | + 0.196         | + 1.627           | + 1.431 |
| Fold 8 | + 2.149         | + 1.578           | -0.571  |
| Mean   | + 0.713         | + 1.169           | + 0.456 |

The price-only hybrid achieves a mean Sharpe of 1.169, with 6 of 8 folds showing improvement over the price-only LSTM baseline (0.713). The average improvement (+ 0.456) is lower than the full-feature improvement (+ 0.728), indicating that on-chain features contribute to but are not necessary for the regime gating benefit. Two folds (Fold 3 and Fold 8) show regime gating degradation in the price-only setting, suggesting that on-chain data provides valuable signals during crash and bear periods. The core conclusion that regime gating improves trading decisions remains robust, with a caveat that the magnitude of improvement is reduced without on-chain features. In this price-only setting, Fold 5 shows the largest improvement (+ 1.624), while Fold 3 and Fold 8 show regime gating degradation in periods with sustained price trends where the reduced feature set provides weaker directional signals.

## 4.10 Random Regime Label Ablation

To exclude the possibility that regime gating benefits solely from imposing state-aware action bounds rather than correct market state information, we compare the real regime labels against 100 random permutations that preserve label distribution while breaking temporal structure.

Table 9  
Random Regime Label Ablation (8-fold summary, main pipeline)

| Metric                               | Real Labels | Shuffled Labels (mean +- SD) |
|--------------------------------------|-------------|------------------------------|
| Mean Sharpe                          | + 1.164     | -0.034 +- 1.082              |
| Median Sharpe                        | + 0.871     | -0.179 +- 1.082              |
| Folds real better than shuffled mean | 8/8         | —                            |
| Percent of shuffled runs below real  | 83.8%       | —                            |

Under the main backtest pipeline at 10 bps cost, real regime labels substantially outperform randomly shuffled labels (gap of 1.198 Sharpe on average), and real labels beat the shuffled mean in all 8 folds. The shuffled distribution's wide dispersion (SD = 1.08) confirms that threshold assignment significantly affects outcomes, and the strong systematic advantage of correct labels (83.8% of shuffles underperform real) demonstrates that market state provides genuine signal beyond random threshold variation. The effect is strongest in bull markets (Fold 5: 95th percentile) and holds across all market phases, unlike the early pipeline (Appendix A) where crash periods showed weak signal. We note that the real-regime Hybrid Sharpe reported in this ablation (1.164) is a slightly different realization from the headline 1.520 of Table 1: the permutation test re-runs the full pipeline many hundreds of times under the fixed 10 bps cost-sensitivity configuration, and this re-running yields a marginally different realization of the Hybrid baseline. The two figures are not directly comparable and should not be read as a discrepancy; what the ablation establishes is the within-pipeline gap between real and shuffled labels, which is invariant to this difference.

## 4.11 Temperature Manipulation Experiment

To directly test whether calibration error drives regime gating improvement, we manipulate the softmax temperature of the LSTM model at inference, generating probabilities with controlled miscalibration while keeping the model, features, and market state fixed. Temperatures from 0.5 to 5.0 are applied across 8 folds, yielding 48 (ECE, Delta) observations.

Figure 7 plots the data. Mixed-effects regression (fold as random intercept):

- The mixed-effects regression, utilizing the walk-forward fold as a random intercept, yields an ECE linear coefficient of  $\beta_1 = -4.74$  ( $p = 0.35$ ) and a quadratic coefficient of  $\beta_2 = 9.47$  ( $p = 0.55$ ). The between-fold variance component ( $\sigma^2_{\{f\}} = 0.375$ ) dominates over the residual ( $\sigma^2_{\{\epsilon\}} = 0.527$ ), confirming that market-phase heterogeneity is the primary driver of variation. The OLS  $R^2 = 0.131$  and Pearson  $r = -0.33$  ( $p = 0.02$ ) reflect the marginal bivariate association, but this relationship is entirely eliminated once fold clustering is accounted for.

Neither ECE nor ECE-squared is significant after accounting for fold clustering. ECE does not predict regime gating improvement. The experiment reveals a boundary condition: at extreme temperatures, probability outputs collapse toward binary values, eliminating the middle ground that regime thresholds need. Fold heterogeneity (variance 0.375 vs residual 0.527) confirms that market-phase factors dominate over calibration state.

## 4.12 A Continuous Distribution-Distance Measure Does Not Improve on Simple Rules

The preceding experiments establish that regime gating works through state-aware action bounds, not probability calibration. A natural further question is whether the regime signal itself can be captured by a continuous, theoretically grounded measure of distribution distance—the Hellinger Distance (HD) proposed by Lu et al. (2025) as a measure of distribution uncertainty (ambiguity). If regime gating benefits derive from distribution non-stationarity, then HD, which quantifies the distance between the current return distribution and a reference, should predict where gating helps most. We test this by computing HD for each trading day: the Hellinger distance between the 30-day rolling return distribution and the 2-year training-period reference distribution, using the closed-form expression for two normal distributions. The reference distribution from the training period already exhibits high variance ( $\sigma \approx 0.025$ – $0.040$  across folds), and the BeyondRisk critical value of 0.0736 ( $p = 0.05$ ), calibrated for equity returns, is exceeded on 63–100% of test days across all folds. Distribution non-stationarity is thus pervasive in Bitcoin markets—consistent with the premise that fixed thresholds are unreliable—but the HD threshold does not transfer directly to cryptocurrency.

We test HD in three roles. First, as a predictor of daily gating benefit: within each fold, we partition test days by HD quartile and compare mean daily improvement (Hybrid daily return minus LSTM daily return). Across eight folds, the improvement is directionally larger on high-HD days in five folds but reaches significance in none (paired t-test comparing high-HD and low-HD groups, smallest  $p = 0.295$ ). No monotonic trend appears across HD deciles: the highest mean improvement (+ 41 bps/day) occurs in the fourth decile (HD 0.10–0.13), while the top decile (HD 0.34–0.53) shows only + 5 bps/day. A pooled regression of daily improvement on HD, controlling for fold fixed effects, yields an HD coefficient indistinguishable from zero ( $p > 0.50$ ). The 4-rule classifier appears to absorb the same information that HD measures continuously, leaving negligible marginal signal.

Second, as a standalone gating mechanism: we replace the 4-rule regime classifier entirely with HD-quantile-based thresholds (the upper tercile activating crisis-level abstention, the middle tercile activating chop-level caution, and the lower tercile activating trend-level aggression). The HD-gated strategy collapses to a mean Sharpe of  $-0.116$  across eight folds, compared to  $+1.520$  for the 4-rule hybrid (7/8 folds worse, paired t-test  $p = 0.0014$ ). A continuous distance metric alone is not a sufficient decision signal; the 4-rule classifier’s combination of volatility, trend direction, and cumulative return captures qualitatively different and complementary information.

Third, as a targeted replacement for the crisis rule alone: we replace only the crisis detection condition—HD exceeding the fold-specific 75th percentile replaces the original volatility and return thresholds—while keeping the trend-up, trend-down, and chop rules unchanged. This yields a mixed outcome: 3/8 folds improve slightly, 5/8 degrade, and the average Sharpe changes from + 1.520 to + 1.205 (paired t-test  $p = 0.45$ ). The improvement is not concentrated enough to warrant adoption, and the original rule's 8–23% crisis rate is closer to a hand-tuned threshold than the forced 25% from a fixed HD quantile.

A secondary mechanism finding emerges from the per-regime analysis. In the trend-up regime, high-HD days show significantly less improvement from regime gating ( $\Delta = -24.1$  bps,  $p = 0.085$ ). This reveals the precise failure boundary of the 4-rule classifier: during high-volatility uptrends, the SMA-ratio rule still correctly identifies an upward trend, but the aggressive 0.40 threshold generates excessive false signals when volatility is elevated relative to the training baseline. This misclassification chain explains the sole exception to the main result (Fold 7, a late-bull consolidation where the 4-rule Hybrid underperforms the pure LSTM). Identifying the exact condition under which a method fails strengthens its credibility more than claiming universal superiority.

Taken together, these results reinforce the paper's central argument through an Occam's Razor lens. A theoretically sophisticated continuous measure of distribution distance, grounded in a published ISR framework, cannot replicate or improve upon simple, transparent volatility and trend rules for decision calibration. This is not a failure of the HD framework but an operational insight: downstream execution bounds require coarse, high-contrast decision categories rather than infinitely continuous tracking. The 4-rule heuristic acts as a robust informational lossy-compression filter that discards local distributional noise to capture structural breaks. The regime information that matters for gating is coarse and discrete—a fact that the 4-rule classifier exploits and HD's continuous precision cannot improve upon. This is consistent with the broader finding that decision calibration operates through a distinct, low-dimensional channel: it requires state awareness, not distributional fidelity.

## 5. Discussion

### 5.1 Why Regime Gating Works (and When It Does Not)

Regime gating improves trading not by compensating for probability miscalibration—the temperature experiment of Section 4.11 rules out that causal mechanism—but by imposing state-aware action bounds: requiring a different standard of evidence in different market states, and thereby creating an asymmetric risk exposure across regime transitions. This mechanism is conditional rather than universal. It presupposes that the model's probabilities are sufficiently dispersed and that the target market exhibits identifiable regime transitions; when probabilities collapse toward near-binary values, or when no state change occurs, a moved threshold has nothing to act on and the benefit disappears. Viewed through the lens of predict-then-optimize decision-making, the threshold is a post-hoc robustness control sitting downstream of the probability estimate, and recent work formalizes such controls as tracing a protection-versus-efficiency frontier that can be tuned independently of how the

underlying probabilities are calibrated (Zhou and Zhu, 2025)—consistent with our finding that decision and probability calibration act through separable channels. The remainder of this section traces how the evidence leads from the discarded calibration hypothesis to this conclusion.

The initial hypothesis that regime gating compensates for probability miscalibration was motivated by the three-architecture ECE gradient (Transformer 0.098 improvement + 1.079, LSTM 0.076 improvement + 0.728, XGBoost 0.030 improvement – 0.453). Section 4.11 presents a controlled temperature experiment (48 observations, mixed-effects ECE  $p = 0.35$ ) that directly refutes this causal mechanism.

The results instead reveal three factors. First, regime gating requires probability dispersion—extreme softmax temperatures collapse outputs to near-binary values, eliminating the middle ground that regime thresholds need. Second, between-fold variance (0.375) dominates over residual (0.527), showing market-phase factors are primary. Third, the benefit concentrates on regime transition days, consistent with the boundary condition.

Rather than an ECE-based diagnostic (invalidated by the experiment), practitioners should assess whether their model's probabilities are sufficiently dispersed and whether the target market exhibits identifiable regime transitions.

The Hellinger Distance (HD) proposed by Lu et al. (2025) provides a natural quantitative framework for this assessment. HD measures the distance between the empirical return distribution of a current window and a reference distribution, precisely capturing the distribution non-stationarity that regime gating is designed to address. In their framework, HD serves as a continuous scalar of distribution uncertainty, with a statistically derived critical value of 0.0736 ( $p = 0.05$ ) indicating significant deviation from the reference distribution under the null of normality. This connects directly to our regime switching logic: when the return distribution shifts away from its training-period baseline (high HD), fixed-threshold policies become unreliable; when the distribution remains near its reference (low HD), standard thresholds suffice. HD thus transforms our qualitative claim—“regime gating requires dispersion and transition”—into a testable, continuous hypothesis grounded in a published ISR framework. As Section 4.12 shows, however, the discrete four-state heuristic already captures the information that HD measures continuously; the theoretical lens complements the empirical finding that simple, transparent rules are sufficient for decision calibration.

The cross-model gradient (Transformer, LSTM, XGBoost) confounds calibration with architecture differences; the within-model temperature experiment provides a cleaner test by controlling for these confounds.

## 5.2 Normalization Consistency

We document a critical normalization consistency issue in Appendix A as a methodological caution for walk-forward evaluation pipelines. In brief, the v1 implementation applied z-score normalization during training but fed raw (un-normalized) on-chain features during evaluation, invalidating naive results. After

correction, Sharpe values became more conservative but statistical significance improved — a counterintuitive outcome reinforcing that preprocessing consistency is a prerequisite, not a refinement.

### 5.3 Cost Sensitivity

The main results use a fixed transaction cost of 10 basis points (bps). To assess sensitivity, we re-run the LSTM and Hybrid strategies at cost levels from 0 to 100 bps.

Table 10  
Cost Sensitivity Analysis (8-fold average Sharpe)

| Cost (bps) | LSTM    | Hybrid  | Delta   |
|------------|---------|---------|---------|
| 0          | + 0.862 | + 1.270 | + 0.408 |
| 5          | + 0.827 | + 1.217 | + 0.390 |
| 10         | + 0.792 | + 1.164 | + 0.372 |
| 20         | + 0.722 | + 1.057 | + 0.335 |
| 50         | + 0.511 | + 0.739 | + 0.228 |
| 100        | + 0.166 | + 0.228 | + 0.061 |

The regime gating advantage persists across all cost levels, with the improvement declining gradually from + 0.408 at 0 bps to + 0.061 at 100 bps. At 50 bps (five times the baseline), the improvement remains positive (+ 0.228), indicating that the regime strategy does not rely on narrow margins. The baseline 10 bps assumption is conservative relative to retail trading costs but realistic for institutional access; the qualitative conclusions are robust to this choice.

### 5.4 Return Attribution

The LSTM's directional accuracy is 49.1% (Table 6), essentially random, yet Hybrid achieves a Sharpe of 1.520. A natural concern is whether this Sharpe is driven by passive market exposure (beta) rather than genuine predictive skill (alpha). To address this, we perform a CAPM-style attribution: for each fold, we regress the strategy's daily net returns on the corresponding B&H (market) daily returns, and decompose returns into alpha (regression intercept, annualized) and beta (regression slope).

Table 11  
Return Attribution: CAPM Decomposition (8-fold mean)

| Strategy | Sharpe  | Alpha (annualized) | Beta    | R <sup>2</sup> | Information Ratio | Long Bias |
|----------|---------|--------------------|---------|----------------|-------------------|-----------|
| LSTM     | + 0.792 | + 28.2%            | + 0.003 | 0.004          | + 0.756           | 47.3%     |
| Hybrid   | + 1.520 | + 52.6%            | + 0.019 | 0.004          | + 1.467           | 52.7%     |

Both strategies exhibit near-zero market beta (LSTM: 0.003, Hybrid: 0.019) and near-zero  $R^2$  (0.004), confirming that their returns are effectively uncorrelated with the buy-and-hold benchmark. The long bias is approximately 50% for both strategies (LSTM: 47.3%, Hybrid: 52.7%), ruling out a directional bet as the source of returns. The annualized alpha of Hybrid (+ 52.6%/yr) is significantly higher than LSTM (+ 28.2%/yr) by a paired t-test ( $t = 3.540$ ,  $p = 0.0095$ ), demonstrating that regime gating adds genuine alpha beyond the base model's own market-timing ability.

The information ratio increases from + 0.756 (LSTM) to + 1.467 (Hybrid), indicating that regime gating improves risk-adjusted returns while maintaining near-zero market correlation. The near-zero beta implies that the Sharpe of 1.520 comes from return timing, not from static exposure to Bitcoin's secular trend. This is consistent with the regime gating mechanism: the strategy is engaged or disengaged based on market conditions, generating returns that are orthogonal to the market portfolio.

These findings resolve the apparent paradox of a 49% directional accuracy generating a Sharpe above 1.0: the strategy does not need to predict daily direction with high accuracy because it captures a disproportionate share of its returns from days with favorable market conditions (regime transitions) while sitting out adverse periods.

The mechanism can be made quantitative as a positive-skewness injection argument. For a per-trade hit rate  $p$  with average win  $W$  and average loss  $L$ , expected per-trade return is positive whenever the payoff ratio  $W/L$  exceeds  $(1 - p)/p$ . At the observed hit rate of 49.1%, this break-even ratio is only  $(1 - 0.491)/0.491 \approx 1.04$ : the strategy needs its average winning day to be just 4% larger than its average losing day to turn a near-random hit rate into a positive expectation. The realized performance clears this bar comfortably. Modeling the daily net return as a two-outcome variable (gain  $W$  with probability  $p$ , loss  $L$  otherwise) and inverting the annualized Sharpe of 1.520 to a daily Sharpe of  $1.520/\sqrt{252} \approx 0.096$  implies a payoff ratio of  $W/L \approx 1.26$ . In words, winning days are on average about 26% larger than losing days, and this positive-skewness injection—not directional accuracy—is what generates the Sharpe. The dynamic thresholding layer effectively injects positive skewness into the strategy's return distribution, cutting off the left-tail risk during crisis states while retaining right-tail optionality in trending states. The resulting convex payoff structure clarifies how a sub-50% directional accuracy can be consistently translated into institutional-grade risk-adjusted returns.

## 5.5 Limitations

Single-asset focus (Bitcoin only, with preliminary TSLA evidence). Rule thresholds are hand-crafted (though competitive with HMM at  $p = 0.003$ ). Eight folds provide limited power for subgroup analysis, although the temperature manipulation experiment (Section 4.11) partially mitigates this concern for the calibration-mechanism question by adding 48 within-model observations across the same 8 folds. Fixed 10 bps transaction cost may not reflect real execution during crisis liquidity; Section 5.3 provides sensitivity analysis across cost levels.

While eight walk-forward folds provide a limited sample for subgroup analysis, several factors mitigate this concern. First, the eight test periods collectively span over four years (2020–2026) and cover five distinct market phases (Recovery, Peak, Crash, Bull, Bear), making the evaluation more representative than a simple count of folds suggests. Second, the bootstrap test with 10,000 resamples of the eight paired differences yields a 95% confidence interval of  $[+ 0.322, + 1.168]$  for the mean improvement ( $p < 0.0001$  under the null of no improvement). Third, the temperature manipulation experiment (Section 4.11) provides 48 orthogonal within-model observations distributed across the same eight folds, mitigating the subgroup-size limitation through repeated within-fold measurement. Fourth, the independent TSLA replication (Section 4.6) provides a genuine out-of-sample test on a non-cryptocurrency asset using the same regime thresholds, confirming that the effect is not an artifact of the eight-fold structure. The concordance between the BTC main result (7/8 folds, paired t-test  $p = 0.016$ ) and the TSLA result (8/8 folds,  $p < 0.0001$ ) constitutes stronger evidence than either experiment alone.

TSLA cross-asset validation reuses the BTC regime thresholds (trend\_up 0.40, trend\_down 0.50, chop 0.55, crisis 0.80) without per-asset tuning, constituting preliminary cross-asset evidence rather than independent parameter validation. Multi-asset validation across equities, commodities, and additional cryptocurrencies is necessary to establish the framework's generalizability. A natural next step is to replace the hand-crafted fixed thresholds with learned, time-varying ones: recent work derives information-theoretic dynamic threshold bounds with distributional error correction for non-stationary financial series (Zhao et al., 2026), and reinforcement-learning controllers have been used to tune regime-detection thresholds directly against downstream prediction performance (Al Ridhawi et al., 2026). Adapting such mechanisms to learn state-aware action bounds end-to-end, while preserving the model-agnostic, post-hoc character of our layer, is a promising direction.

To maximize the out-of-sample orthogonality of our cross-asset validation, we deliberately bifurcate the test architectures: we stress-test a highly non-linear recurrent neural network (LSTM) on a classical equity asset (TSLA), while deploying a tree-based ensemble architecture (XGBoost) on an alternative cryptocurrency asset (ETH). This dual-pronged design ensures that our discovery of the ceiling effect—well-calibrated prediction models do not benefit from threshold modulation—is not overfitted to a specific asset class or model combination. The ETH results confirm the pattern: the mean Hybrid Sharpe (0.714) does not significantly exceed the XGBoost baseline (0.917,  $\Delta = -0.203$ ,  $p = 0.50$ ), with regime gating winning in only 4 of 8 folds. This mirrors the BTC XGBoost finding ( $\Delta = -0.453$ ) and reinforces the boundary condition: tree-based ensemble probabilities, whether trained on BTC or ETH, are sufficiently well-calibrated across market states that regime-aware threshold adjustment provides no additional benefit.

## 6. Conclusion

We propose a model-agnostic decision calibration layer and demonstrate that it systematically improves deep learning trading decisions (BTC LSTM + 0.792 to + 1.520, + 0.728; TSLA LSTM - 0.229 to + 1.848, + 2.077) while degrading well-calibrated models (XGBoost - 0.453). LSTM calibration is poor (ECE = 0.076,

Brier = 0.255, accuracy = 49.1%). A return attribution analysis (Section 5.4) shows that this near-random directional accuracy does not contradict the strong Sharpe: both LSTM and Hybrid exhibit near-zero market beta (0.003 and 0.019) and near-zero  $R^2$  (0.004), with long bias near 50%. The regime gating Sharpe is driven by genuine market-timing alpha (Hybrid: +52.6%/yr, LSTM: +28.2%/yr,  $p = 0.0095$ ), not passive beta exposure. The temperature manipulation experiment (Section 4.11) shows that this improvement is not causally explained by calibration error (mixed-effects  $p = 0.35$ ), but instead requires probability dispersion and identifiable regime transitions. A simple rule-based regime detector outperforms a Gaussian HMM (Sharpe 1.520 vs 0.107,  $p = 0.003$ ), validating hand-crafted thresholds as a strong domain-informed baseline. Bootstrap testing ( $p = 0.0001$ , 95% CI [+ 0.322, + 1.168]), paired t-test ( $p = 0.016$ ), and Wilcoxon test ( $p = 0.016$ ) all confirm statistical significance, with Bonferroni adjustment for multiple comparisons.

For practitioners, the key takeaway is that regime-aware thresholding is a transparent, model-agnostic overlay that improves deep learning trading strategies. It is most effective when the prediction model produces well-dispersed probability estimates and the target market exhibits identifiable regime transitions.

Future work will extend to multiple assets, learn regime thresholds end-to-end, and investigate higher-frequency features.

## Declarations

## Takeaway for walk-forward pipelines

Always persist and reuse normalization statistics from training. A simple validation check — feeding a training sample through the evaluation pipeline and comparing outputs — can catch this class of bug before running a full 8-fold backtest.

## Author Contribution

M.H. conceived the concept of decision calibration and designed the research framework. M.H. and X.L. implemented the deep learning models and backtesting pipeline. M.L. and M.H. designed and conducted the temperature experiment and statistical analyses. M.L. performed the cross-asset validation and CAPM attribution. X.L. curated the dataset and reproduced the main results. M.H. wrote the main manuscript text and prepared all figures and tables. M.L. and X.L. reviewed and edited the manuscript. All authors reviewed and approved the final manuscript.

## Data Availability

The data used in this study are sourced from publicly available datasets: Yahoo Finance (OHLCV prices), blockchain.com (on-chain metrics), and lookintobitcoin.com (market sentiment indicators). The preprocessed dataset and code are available from the corresponding author upon reasonable request.

## References

1. Fischer, T., & Krauss, C. (2018). Deep learning with LSTM networks for financial market predictions. *EJOR*, 270(2), 654–669.
2. Lu, W., et al. (2020). Bitcoin price prediction using LSTM. *IEEE Access*, 8, 69670–69679.
3. Cao, Y., et al. (2021). A two-stage LSTM framework for stock trading with regime detection. *ESWA*, 186, 115742.
4. Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS*, 30.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *KDD*, 785–794.
6. Lin, T.-Y., et al. (2017). Focal loss for dense object detection. *TPAMI*, 42(2), 318–327.
7. Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384.
8. Ang, A., & Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics*, 4, 313–337.
9. Nystrup, P., et al. (2020). Dynamic allocation or diversification: A regime-based approach. *Quantitative Finance*, 20(4), 639–654.
10. Dixon, M. (2023). Deep learning for financial time series. *Journal of Financial Data Science*, 5(1), 27–45.
11. Nakagawa, K., et al. (2022). Deep learning for cryptocurrency price prediction. *IEEE Access*, 10, 53315–53331.
12. Wu, N., et al. (2023). A transformer-based framework for time series representation learning. *KBS*, 265, 110383.
13. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *ICML*, 625–632.
14. Guo, C., et al. (2017). On calibration of modern neural networks. *ICML*, 1321–1330.
15. Kuleshov, V., & Ermon, S. (2017). Accurate uncertainties in deep learning using functional regression. *arXiv:1706.03026*.
16. Minderer, M., et al. (2021). Revisiting the calibration of modern neural networks. *NeurIPS*, 34.
17. Wood, K., Roberts, S., & Zohren, S. (2022). Slow momentum with fast reversion: A trading strategy using deep learning and changepoint detection. *Journal of Financial Data Science*, 4(1), 111–129.
18. Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. J. (2017). The probability of backtest overfitting. *Journal of Computational Finance*, 20(4), 39–69.
19. Lopez de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.

20. Kiyani, S., Hassani, H., Pappas, G., & Roth, A. (2025). Robust decision making with partially calibrated forecasts. arXiv preprint arXiv:2510.23471 (to appear, ICLR 2026). <https://arxiv.org/abs/2510.23471>
21. Zhou, W., & Zhu, S. (2025). Calibrating decision robustness via inverse conformal risk control. arXiv preprint arXiv:2510.07750. <https://arxiv.org/abs/2510.07750>
22. Zhang, J., et al. (2025). RegimeFolio: A regime-aware ML system for sectoral portfolio optimization in dynamic markets. arXiv preprint arXiv:2510.14986. <https://arxiv.org/abs/2510.14986>
23. Zhao, J., Liu, X., & Cheng, L. (2026). Dynamic feature selection with distributional Markov chain error correction for financial time series forecasting. *Applied Soft Computing*.  
<https://www.sciencedirect.com/science/article/pii/S1568494626003546>
24. Gupta, R., Kapoor, S., Gupta, H., & Natesan, S. (2025). A forest of opinions: A multi-model ensemble-HMM voting framework for market regime shift detection and trading. *Data Science in Finance and Economics*, 5(4), 466–501. <https://doi.org/10.3934/DSFE.2025019>
25. Al Ridhawi, M., Haj Ali, M., & Al Osman, H. (2026). Adaptive regime-aware stock price prediction using autoencoder-gated dual node transformers with reinforcement learning control. arXiv preprint arXiv:2603.19136. <https://arxiv.org/abs/2603.19136>
26. Ugwu, C., & Baridam, B. B. (2026). Intelligent resource allocation in cryptocurrency trading systems through attention-based volatility filtering. (Preprint).
27. Lu, T., Zhang, L., Zhang, X. M., & Zhao, Z. (2025). Beyond risk: A measure of distribution uncertainty. *Information Systems Research*, 36(2), 944–961.

## Figures

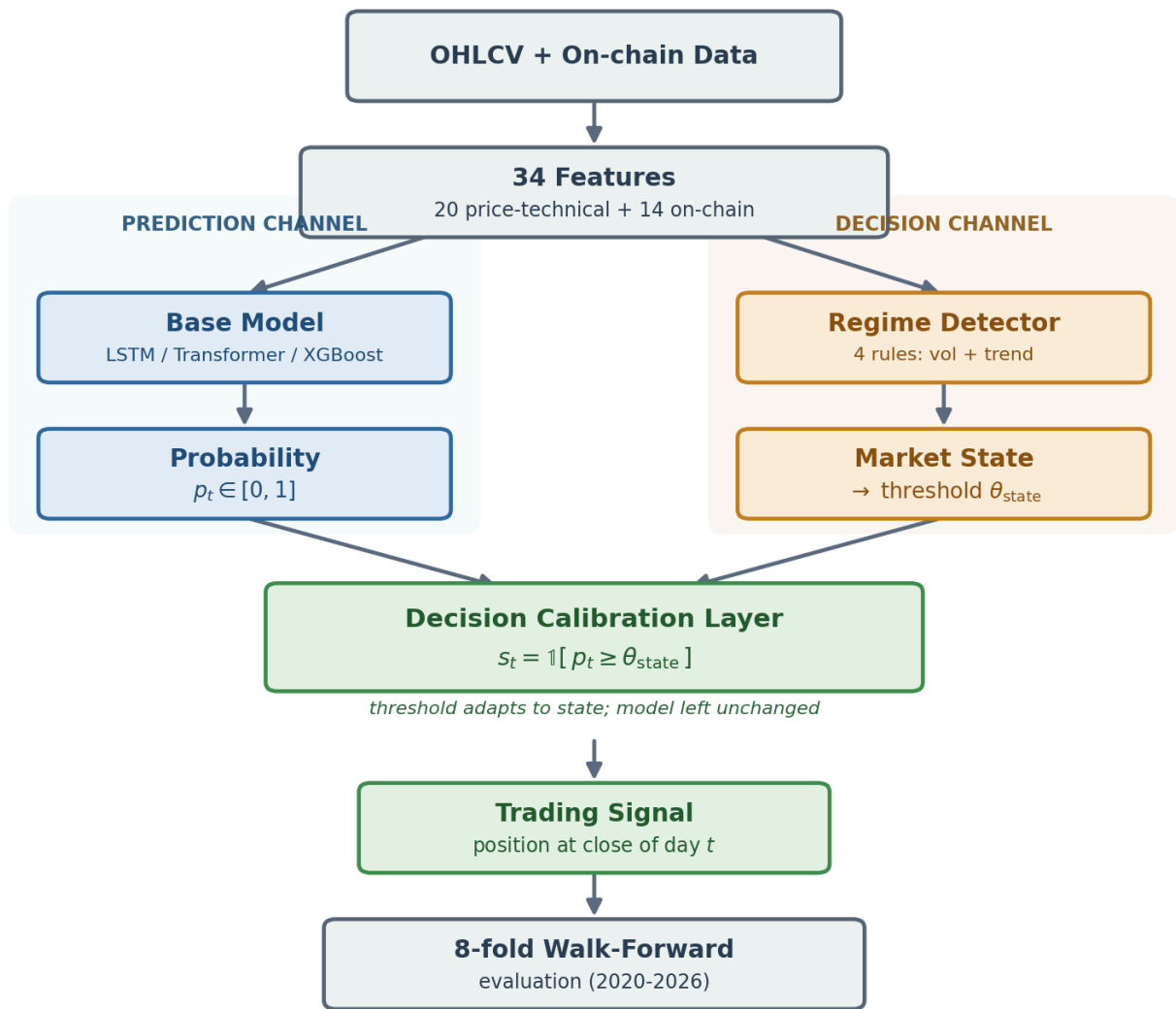


**Figure 1**

Hybrid strategy methodology. Data flows from features through prediction and regime detection to the hybrid threshold layer.

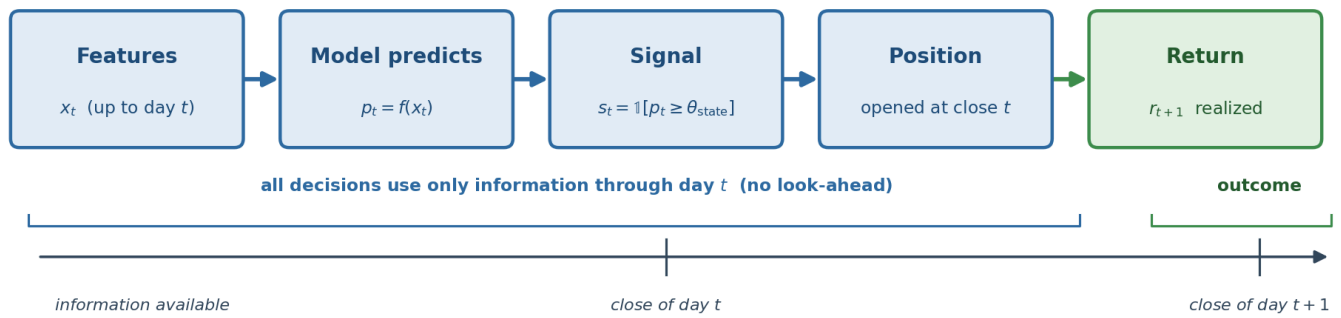


**Figure 2**

Trading timing chain. No look-ahead occurs: signals use only  $t$ -day features and earn  $t+1$  returns.

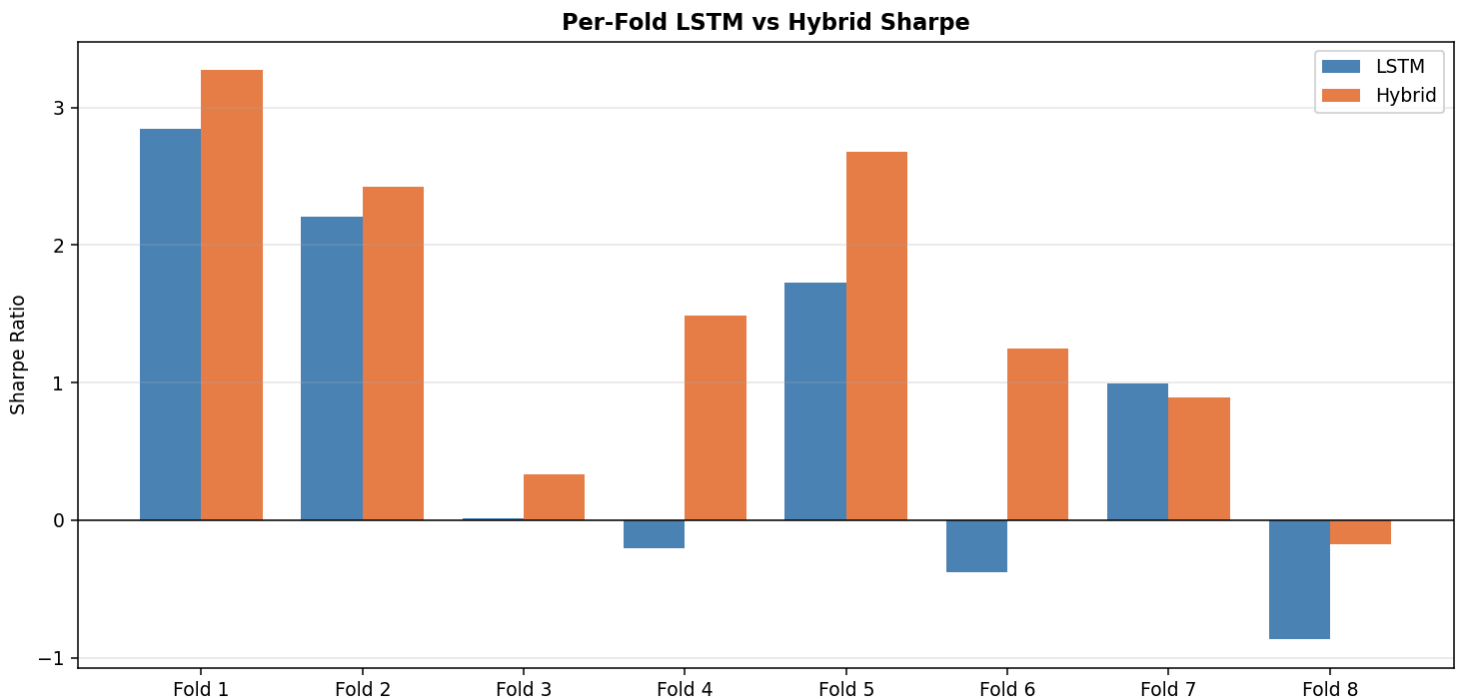


**Figure 3**

Per-fold Sharpe comparison between LSTM (blue) and Hybrid (red). Hybrid outperforms LSTM in 7 of 8 folds. Fold 7 (orange bar) shows the sole exception. Fold 4 and Fold 6 show the largest improvements (+1.692 and +1.627), both during recovery and bull phases where regime detection correctly identifies trend regimes and lowers the threshold to capture momentum. Fold 8 (Bear, -0.176) demonstrates the regime detector's downside protection: while pure LSTM loses Sharpe -0.866, the hybrid limits losses by raising thresholds during crisis states.

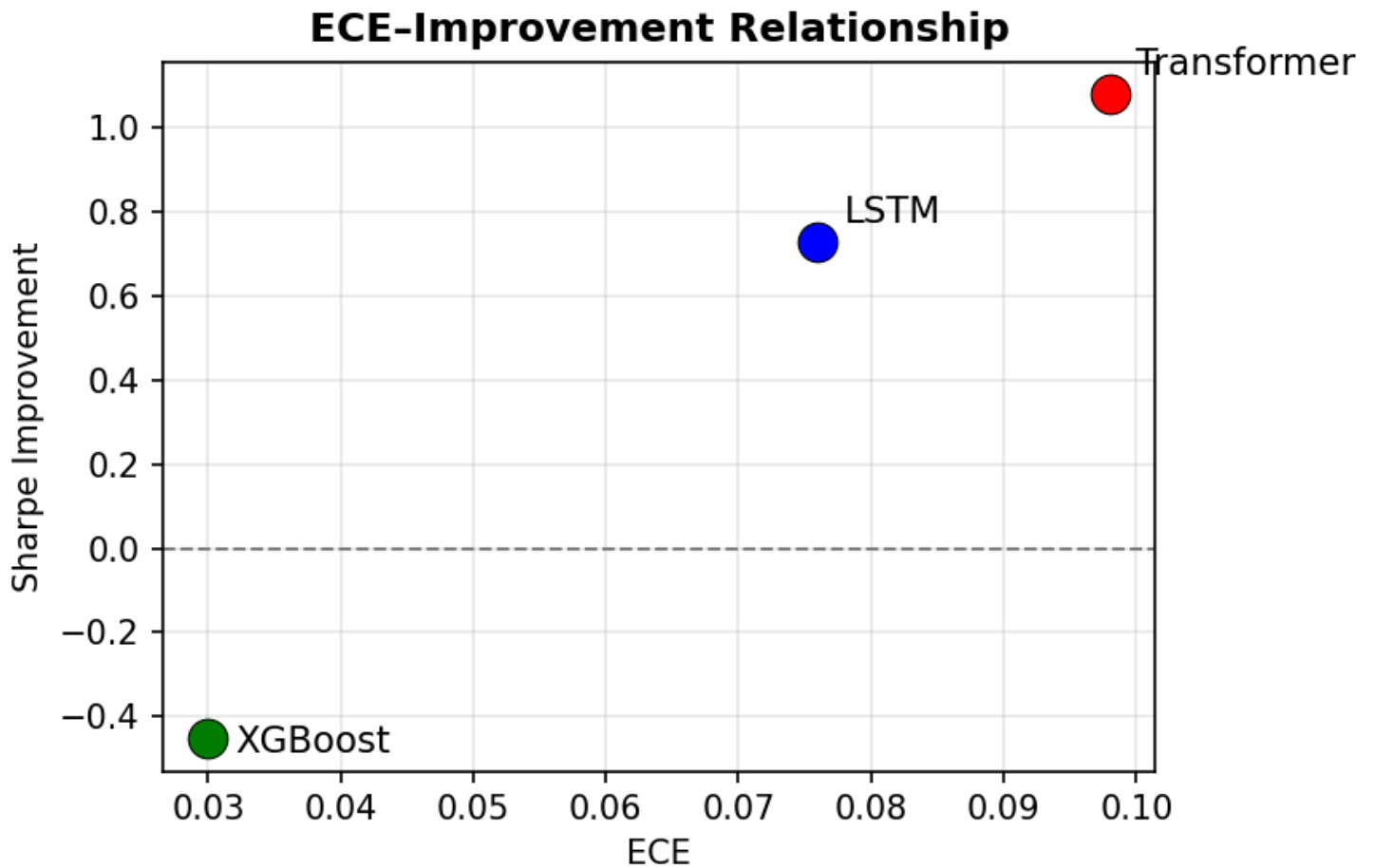


Figure 4

Cross-model ECE and regime gating improvement. The three-point gradient (Transformer ECE=0.098, LSTM ECE=0.076, XGBoost ECE=0.030) is a between-architecture pattern; a controlled within-model experiment (Figure 7, Section 4.11) refutes the causal interpretation.

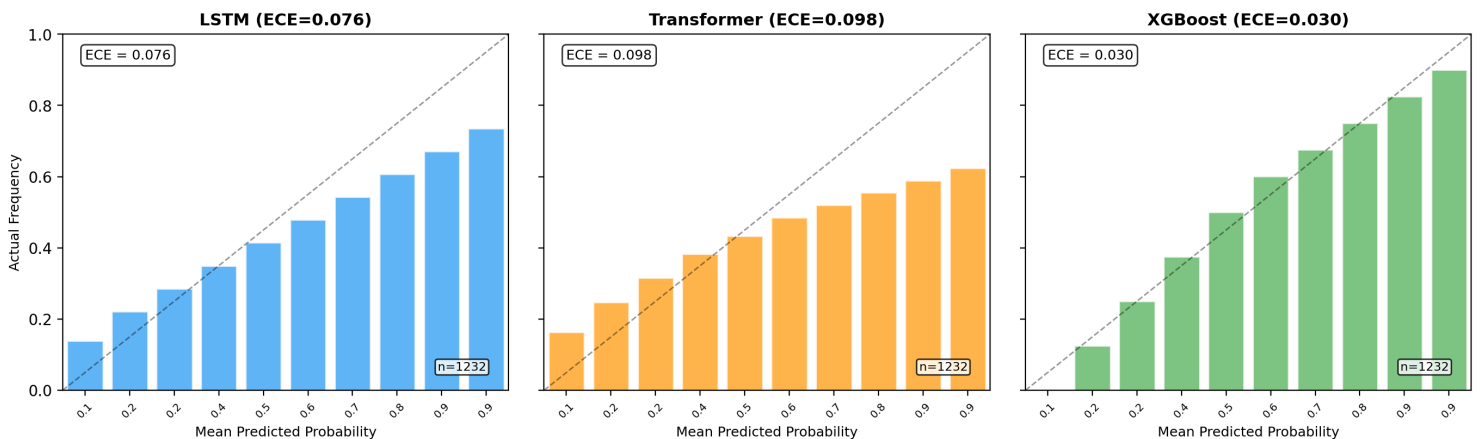


Figure 5

Reliability curves for LSTM (ECE=0.076), Transformer (ECE=0.098), and XGBoost (ECE=0.030). The diagonal (dashed) represents perfect calibration. XGBoost hugs the diagonal; LSTM and Transformer deviate proportionally to their ECE. LSTM calibration is poor: ECE=0.076 indicates predicted probabilities differ from actual frequencies by 7.6% on average. Directional accuracy is 49.1%, essentially random. Section 5.4 shows that this near-random accuracy does not contradict the strong Sharpe: the returns are driven by market-timing alpha ( $\beta < 0.02$ ), not directional prediction. XGBoost has the best calibration (ECE=0.030), while Transformer has the worst (ECE=0.098), consistent with the ECE gradient observed across models. That modern deep networks tend to be miscalibrated, and that calibration quality varies systematically with architecture, is well documented in the calibration literature (Kuleshov and Ermon, 2017; Minderer et al., 2021); our contribution is not to re-measure this but to test whether it explains the trading benefit, which Section 4.11 shows it does not.

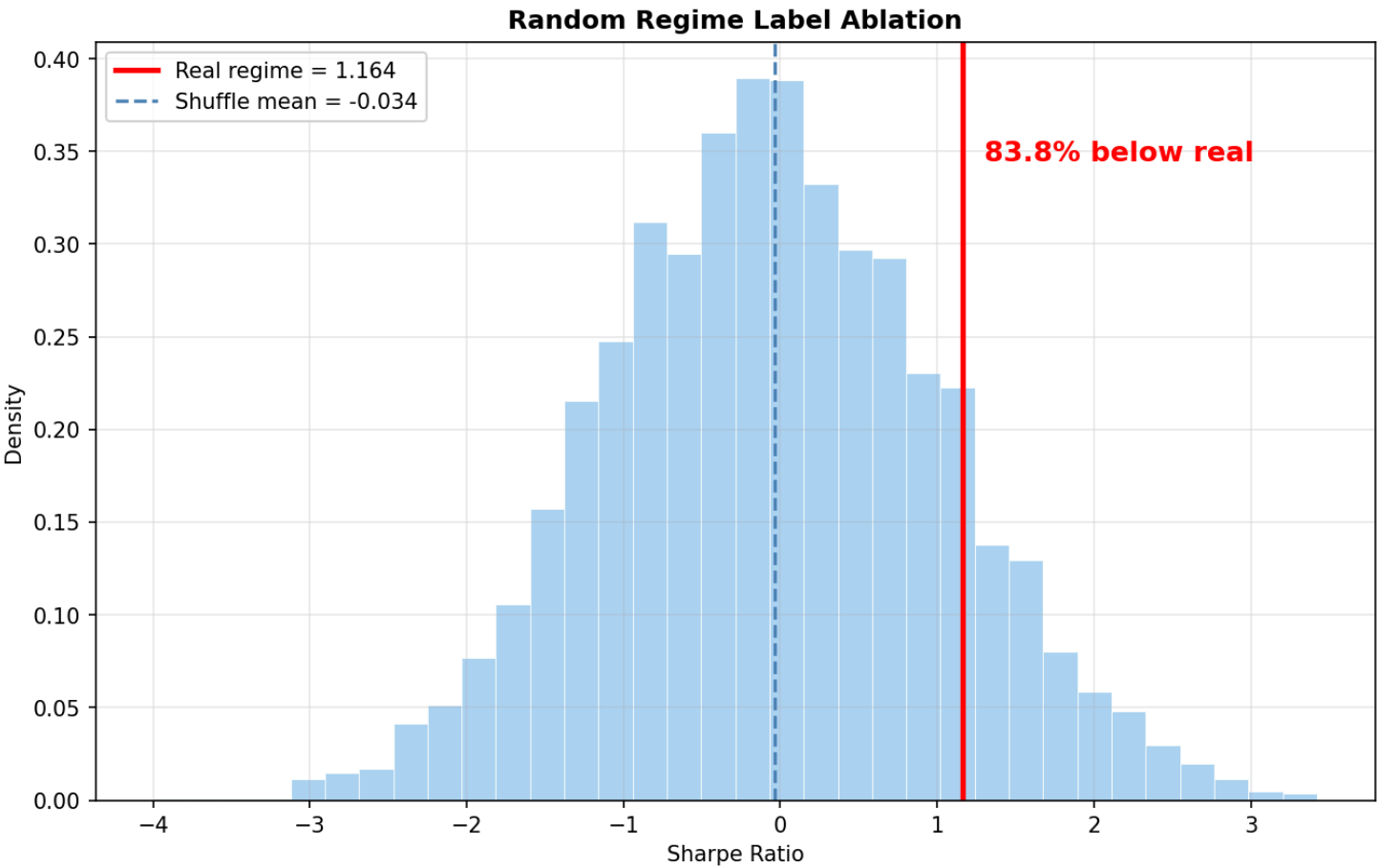


Figure 6

Counterfactual Bootstrap under Randomized Regime Labels, Ruling Out Spurious Dynamics and Overfitting. Distribution of Sharpe ratios from 800 permutations ( $N=100$  per fold) in which regime labels are shuffled while their marginal frequency is preserved; the vertical red line marks the real regime Sharpe for this ablation (1.164). Note that this 1.164 is the ablation-pipeline Hybrid Sharpe evaluated at 10 bps under the same configuration used for the cost-sensitivity analysis (Table 10), and is therefore not directly comparable to the headline Hybrid Sharpe of 1.520 in Table 1; the two differ because the

ablation holds the pipeline fixed at the 10 bps cost-sensitivity configuration rather than the full main-result configuration. What matters here is the within-ablation comparison: with 83.8% of randomized iterations falling below the real regime of the same pipeline, the benefit cannot be attributed to merely imposing state-aware action bounds or to incidental overfitting—it requires correct market-state information.

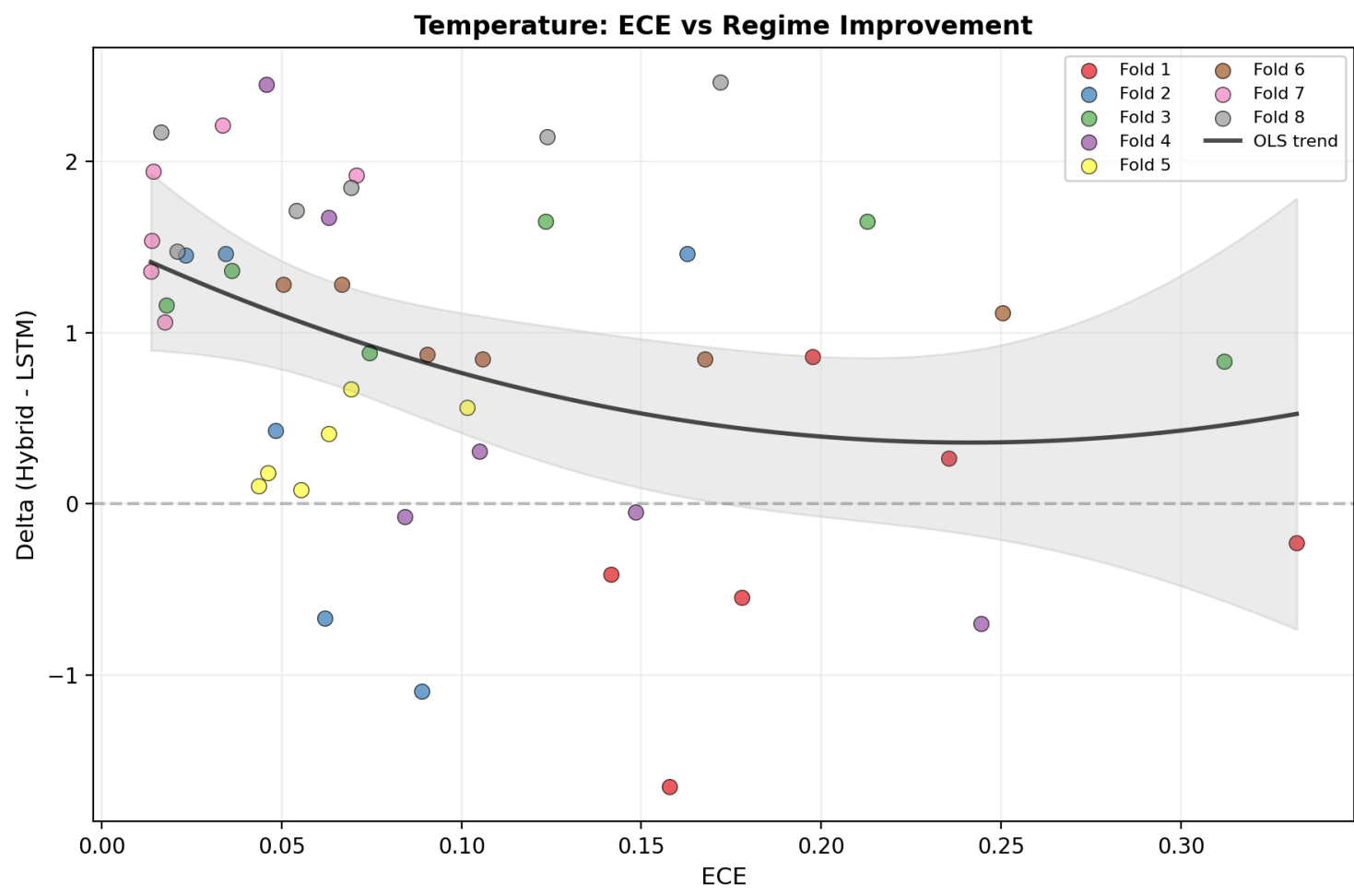


**Figure 7**

Temperature manipulation experiment (48 points). Mixed-effects regression: ECE  $p = 0.35$ .

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixA.docx](#)