Transactions on Quantitative Finance and Beyond

www.tqfb.reapress.com

Trans. Quant. Fin. Bey. Vol. 3, No. 2 (2026) 111–127.

Paper Type: Original Article

Regime Survival Forecasting for Adaptive Execution: Beyond Fixed Aggregation Windows

Satish Garg

Independent Researcher, India; satish.garg08@yahoo.com.

Citation:

Received: 07 August 2025 *Revised: 24 September 2025* *Accepted: 02 January 2026*	Garg, S. (2026). Regime survival forecasting for adaptive execution: Beyond fixed aggregation windows. *Transactions on Quantitative Finance and Beyond, 3*(2), 111-127.

Abstract

We test whether Weibull Accelerated Failure Time (AFT) survival models can replace the fixed aggregation window in Hidden Markov Model (HMM)-based regime-aware execution with a per-instance prediction of remaining bearish regime duration. Fitting models to 54 bearish regime instances across eight asset classes (2020–2024), we find the extension fails on three structural grounds. First, concordance indices of 0.20–0.39 confirm that HMM-derived covariates, posterior entropy and stay probability at regime start, carry no duration-predictive information under parametric AFT models. Second, a without-replacement subsampling simulation shows the concordance index remains flat at approximately 0.48 from $n = 4$ to $n = 45$ training instances, ruling out data quantity as the cause. Third, Weibull shape parameters below 1.0 in every asset class produce decreasing-hazard distributions whose mean durations structurally exceed the window cap, collapsing 60–89% of predictions to boundary values regardless of covariate values. There are no genuine adaptive wins over the fixed ten-day aggregation window established in prior work. Together, Papers I-III characterize the limits of HMM-based regime awareness in execution: regime signals require multi-day aggregation (Paper II), learned policies cannot exploit them reliably (Paper I), and adaptive window calibration via parametric survival models is not achievable with HMM posteriors as duration predictors (Paper III). These findings close a research trilogy with a complete empirical characterization of what HMM-based regime awareness can and cannot achieve in optimal trade execution.

Keywords: Survival analysis, regime detection, Trade execution, Hidden markov model, Weibull AFT, Adaptive aggregation.

1| Introduction

Optimal trade execution is complicated by market regimes. A skilled trader executing a large buy order behaves

Corresponding Author: satish.garg08@yahoo.com

https://doi.org/10.22105/tqfb.v3i2.85

differently depending on whether prices are trending upward, downward, or oscillating: urgency is rewarded in bullish periods, while patience captures favorable prices in bearish ones. Hidden Markov Models (HMMs) provide a natural framework for detecting these regime states, and HMM-based uncertainty signals have been proposed as filters for regime-aware execution systems. Two empirical questions motivate this trilogy of papers. First, can HMM uncertainty signals predict when the regime-aware strategy outperforms a regime-blind baseline? Second, if those signals are informative, can we exploit that information to improve execution in practice?

Papers I and II addressed these questions in sequence. Paper I [6] showed that flat reinforcement learning (RL) agents cannot reliably exploit regime signals even when the true regime label is provided in the state space, identifying the failure as structural, a property of the policy gradient optimization landscape. Paper II [7] then asked whether hand-crafted HMM uncertainty signals, rather than learned policies, could at least predict execution quality. The answer was positive, but conditional: at daily resolution, signals are largely uninformative, while aggregating over $W = 10$ trading days reveals significant predictive power for the iShares Russell 2000 ETF (IWM) entropy ($\rho = -0.411$, $p < 0.001$) and Bitcoin-USD (BTC-USD) stay probability ($\rho = -0.204$, $p < 0.001$). The temporal threshold aligns broadly with empirical mean regime durations.

This finding raises a natural follow-on question: if the appropriate aggregation window depends on regime duration, can we predict regime duration at the start of each bearish instance and use that prediction to set W adaptively? A survival model is the appropriate tool, as it estimates the remaining duration of an ongoing event (here, a bearish regime) from covariates observed at event start. Adaptive windows calibrated to predicted regime duration would replace the fixed $W = 10$ with a per-instance estimate, potentially recovering prediction power that a single fixed window loses on instances with significantly shorter or longer-than-average durations.

We test this extension rigorously. We fit Weibull AFT models to 54 bearish regime instances extracted from five years of daily data across eight asset classes (2020–2024 train; 2023–2024 test). We derive per-day adaptive windows from the survival model predictions and compare the resulting execution signal against fixed windows from Paper II. We find that the adaptive approach fails, and we characterize the failure mechanistically with three findings. First, all survival models achieve concordance indices below 0.50, meaning HMM-derived regime start covariates have no predictive power for subsequent regime duration under parametric AFT models. Second, a sample-size simulation shows this is not a data quantity problem, as the concordance index does not improve from $n = 4$ to $n = 45$ training instances. Third, Weibull shape parameters below 1.0 reveal decreasing-hazard duration distributions in all asset classes, a structural property that makes remaining-duration prediction degenerate.

The contribution of this paper is therefore a negative one in the specific sense: we show that a natural and well-motivated extension of the regime signal framework does not work, and explain precisely why. This closes the trilogy with a complete empirical characterization of what HMM-based regime awareness can and cannot achieve in execution.

Contributions.

I. We conduct the first test of survival-model-based adaptive aggregation windows for regime-aware trade execution, comparing Weibull AFT predictions against fixed windows from prior work across eight asset classes and a genuine out-of-sample test period (2023–2024).
II. We show that HMM-derived regime start covariates (posterior entropy, stay probability) carry no information about subsequent regime duration under parametric AFT models, evidenced by concordance indices of 0.20–0.39 across all assets and pooling strategies.

III. We identify decreasing-hazard duration distributions (Weibull shape < 1.0) across all asset classes as a structural property that makes remaining-duration prediction degenerate, causing adaptive windows to collapse to boundary values.
IV. We provide a sample-size simulation showing that the concordance index remains flat at ≈ 0.48 from $n = 4$ to $n = 45$ training instances, ruling out data quantity as the limiting factor and establishing that the failure is a covariate adequacy problem under parametric models.
V. Together with Papers I and II, we provide a complete empirical characterization of the limits of HMM-based regime awareness in execution.

2| Literature Review

2.1|Optimal Execution and Regime-Aware Strategies

The optimal execution problem arises when an institutional trader must purchase or sell a large quantity of shares before a deadline at minimal cost. Almgren and Chriss [1] derived the classical solution as a mean-variance optimization under linear market impact, producing deterministic execution schedules. Perold [10] introduced implementation shortfall as the standard cost metric. Time-Weighted Average Price (TWAP), splitting an order into equal tranches at equal intervals, serves as the practical regime-blind baseline. All these approaches assume stationary market conditions and cannot adapt to regime shifts that characterize real markets.

Amrouni et al. [2] introduced the CTMSTOU simulation environment, demonstrating that regime-aware hand-coded rules outperform regime-blind strategies. Paper I [6] showed that flat Proximal Policy Optimization (PPO) agents cannot replicate this advantage even with the true regime label in the state space, identifying the failure as structural. Paper II [7] validated that HMM uncertainty signals are empirically predictive of execution quality when aggregated over 3–10 days, establishing the temporal threshold that motivates this study.

2.2|Survival Analysis for Duration Prediction

Survival analysis models the time until an event, accommodating right-censored observations where the event has not yet occurred at the end of the observation window. The Cox proportional hazards model [3] estimates covariate effects on the hazard function semi-parametrically. The Weibull AFT model [4] assumes a parametric form for the survival time distribution, with the AFT parameterization expressing covariate effects as log-time multipliers directly interpretable as duration accelerants. The Weibull shape parameter ρ determines the hazard structure: $\rho < 1$ implies decreasing hazard (the longer a regime persists, the less likely it is to end), $\rho = 1$ implies constant hazard (memoryless), and $\rho > 1$ implies increasing hazard (regimes become more likely to end as they age).

Survival models have been applied extensively in biomedical research and engineering reliability [4] but have not, to our knowledge, been applied to regime duration prediction for execution. The application is natural: a bearish regime is an ongoing event with a start date, and the question of how much longer it will persist maps directly to the survival analysis framework.

2.3|HMM-Based Regime Detection in Trading

Hamilton [8] established the theoretical foundation for Markov-switching models in financial time series. Recent execution systems assume HMM regime signals are informative without testing this empirically: TradeR [12] proposes a two-level hierarchical architecture motivated by regime diversity; EarnHFT [11] trains separate agents per market trend with an HMM-based router; Xu et al. [13] use a mixture-of-experts framework with regime states determining mixing weights. Lopez de Prado [9] provides a comprehensive treatment of overfitting risks in financial machine learning that motivates our out-of-sample validation design. Paper II [7] provides the empirical validation these architectures assume: signals are informative but only at multi-day aggregation horizons of 3–10 days.

3|Methodology

3.1|Data and HMM Fitting

We collect daily open-high-low-close-volume (OHLCV) data for eight assets via `yfinance` over a five-year period (2020-01-01 to 2024-12-31): three equity index ETFs (SPY, QQQ, IWM), two cryptocurrencies (BTC-USD, ETH-USD), one commodity ETF (GLD), one bond ETF (TLT), and one large-cap equity (AAPL). Equity assets yield approximately 1,237 trading days; cryptocurrency assets yield 1,806 calendar days due to continuous trading.

These eight assets are identical to those studied in Papers I and II, a deliberate choice serving two purposes. First, a consistent asset universe enables direct cross-paper comparison: the survival model results here can be related to Paper II's aggregation-window findings and Paper I's RL failure analysis without confounding from different asset coverage. Second, the set was originally constructed to span the widest feasible range of asset-class heterogeneity relevant to institutional execution. SPY, QQQ, and IWM provide large-, mid-, and small-cap equity index exposure with structurally different regime dynamics; BTC-USD and ETH-USD represent cryptocurrency markets with distinct volatility scales and continuous trading; GLD captures a store-of-value commodity whose regimes are driven by risk-off flows rather than earnings cycles; TLT introduces interest-rate regime sensitivity absent from equity assets; and AAPL represents a large-cap single name where idiosyncratic events can dominate regime onset. This breadth ensures that the failure of survival-model-based adaptive windows is characterised across structurally heterogeneous regime environments rather than within a single asset class.

The five-year window is motivated by the survival analysis requirement: with a one-year window as in Paper II, the number of complete bearish regime instances per asset is too small (1–3) for any survival model to be fitted. With five years, we obtain 81 total bearish instances across all assets, providing sufficient pooled samples for class-level models.

We fit a 4-state Gaussian HMM using the identical procedure as Paper II: 40 random initialization seeds, Bayesian Information Criterion (BIC)-based covariance structure selection, and a minimum 3% state occupancy constraint. State labels are assigned via the same two-step scoring procedure (crash, bearish, transitional, bullish). The smoothing procedure suppressing single-day regime switches is retained. All HMM fitting is in-sample over the full five-year window; the out-of-sample contribution in this study comes from the temporal train/test split applied at the survival model level.

In Paper II, HMM fitting used a one-year window; extending to five years here is necessary to accumulate sufficient bearish regime instances for survival model estimation. We note that in-sample fitting over the full five-year window means that regime assignments in the test period (2023–2024) reflect information from the full sample; reported survival model results should therefore be treated as upper bounds on prospective predictive power.

3.2|Regime Collapse and Instance Extraction

Regime collapse. We collapse the four-state HMM sequence to a binary label: crash (state 0) and bearish (state 1) are merged into *bearish*; transitional (state 2) and bullish (state 3) are merged into *non-bearish*. This collapse is motivated by Paper II's finding that the execution edge is concentrated in states 0–2. We acknowledge that crash and bearish states are structurally distinct and that pooling them may introduce heterogeneity; this is noted as a limitation in Section .

Instance extraction. A regime instance is a maximal contiguous run of the collapsed label. Each bearish instance is characterized by its start date, end date, duration in days, and covariates measured at regime start (Section). Two censoring rules are applied. First, any instance still active at the last sample day is right-censored (event indicator = 0). Second, instances exceeding 120 days are treated as censored at that threshold, primarily affecting BTC-USD where a single regime lasted over two years.[1]

[1]The 120-day threshold constitutes administrative censoring applied retrospectively; Section 5.3 confirms the decreasing-hazard finding is robust across alternative thresholds of 90 and 150 days.

Train/test split. Instances with start dates on or before 2022-12-31 form the training set; instances starting on or after 2023-01-01 form the test set. This produces 54 bearish training instances and 27 bearish test instances. *Table 1* summarizes these per asset. BTC-USD contributes no test instances because the 2021–2022 cryptocurrency market downturn produced a single large bearish regime consuming the entire BTC-USD training allocation.

Table 1. Bearish regime instance summary. Train instances: start date $\leq$ 2022-12-31. Test instances: start date $\geq$ 2023-01-01. Med. dur. = median duration across all instances (days). Censored = fraction of instances right-censored.

Asset	Class	n	n_{train}	n_{test}	Med. dur. (d)	Censored
SPY	Equity index	16	10	6	21.5	6.3%
QQQ	Equity index	6	5	1	17.5	16.7%
IWM	Equity index	4	3	1	66.0	25.0%
BTC-USD	Crypto	5	5	0	26.0	20.0%
ETH-USD	Crypto	17	14	3	31.0	17.6%
GLD	Commodity	10	5	5	36.5	30.0%
TLT	Bond	12	4	8	25.0	16.7%
AAPL	Equity	11	8	3	45.0	9.1%
Total		81	54	27		

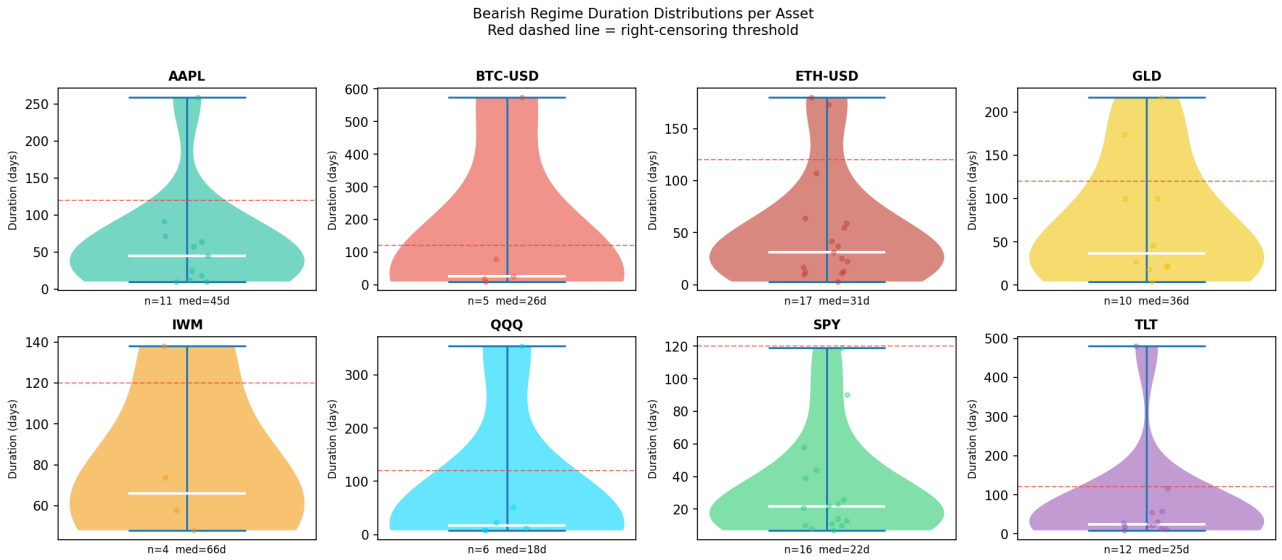

Fig. 1. Bearish regime duration distributions per asset (violin plots). The blue horizontal bar marks the interquartile range; the white bar marks the median. The red dashed line is the 120-day right-censoring threshold. Points show individual regime instances.

3.3|Survival Model Covariates

We use two covariates, both measured on the first day of each bearish regime instance:

I. **Entropy at start** (H_0): HMM posterior entropy on the first day of the regime instance. Lower entropy indicates a cleaner, more well-defined regime transition. The hypothesis is that clearer regime detection at onset predicts more persistent regimes.

II. **Stay probability at start** (SP_0): The transition matrix diagonal element for the current state on the first day of the instance. Higher stay probability indicates the model considers the state inherently persistent.

We exclude *volatility at start* (75% missing values in preliminary runs) and *prior regime duration* (produced extreme coefficients −0.93 log-time units on the AAPL pool, a clear overfitting artifact). Restricting to two covariates is statistically conservative and appropriate given instance counts of 3–19 per asset class.

3.4|Weibull AFT Survival Model

We fit a Weibull Accelerated Failure Time model using the `lifelines` library [5]. The AFT parameterization expresses the log-survival time as a linear function of covariates:

$$\log T_i = \mu + \beta^\top \mathbf{x}_i + \sigma\epsilon_i, \tag{1}$$

where T_i is the duration of instance i, $\mathbf{x}_i$ is the covariate vector at regime start, β is the coefficient vector on the log-time scale, σ is the scale parameter, and ϵ_i follows a Gumbel distribution (yielding Weibull marginal durations). Positive β_j indicates covariate j is associated with longer regime duration. Covariates are standardized via `StandardScaler`. A penalizer of 0.1 (L2 regularization) is applied to stabilize estimation on small samples.

Model hierarchy. A per-asset model is fitted when the asset contributes at least 8 training instances. Otherwise, instances are pooled within asset class. If the pooled class also has fewer than 8 instances, the asset receives a median-fallback predictor. This hierarchy applies to GLD and TLT, which fall back to medians.

The threshold of 8 instances is based on a minimum-observations-per-parameter heuristic: The Weibull AFT model estimates three free parameters (intercept μ, two covariate coefficients β_{H_0} and β_{SP_0}, and scale σ), giving a parameter-to-observation ratio of 1:2 at $n = 8$. Although L2 regularization (penalizer = 0.1) permits estimation below this ratio, the threshold provides a conservative lower bound consistent with standard survival analysis practice for small samples. To assess sensitivity, we re-ran the model hierarchy with alternative thresholds of 6 and 10 instances: at $n_{\min} = 6$, no additional per-asset models become available because the only affected assets (GLD, $n = 5$; TLT, $n = 4$) still fall below the lower threshold; at $n_{\min} = 10$, SPY ($n = 10$) transitions to the pooled equity-index model, yielding a C-index of 0.384 (unchanged from the pooled result in Table 2) and leaving the main conclusions unaffected. The 8-instance threshold is therefore not a critical decision boundary for the reported results.
Predictive evaluation. Concordance index (C-index) measures whether the model correctly ranks pairs of instances by predicted duration. C-index = 0.50 corresponds to chance; C-index > 0.70 is conventionally useful in survival analysis.

3.5|Adaptive Window Generation

For each day t in the test period within a bearish regime instance, the predicted remaining duration is:

$$\hat{R}_t = \max\Big(1,\ \hat{T}_0 - d_t\Big) \tag{2}$$

where $\hat{T}_0$ is the Weibull AFT median survival time predicted from covariates at the start of the current instance, and d_t is the number of days elapsed since the instance start.[2] The adaptive window is:

$$W_t^{\mathrm{adapt}} = \mathrm{clip}\Big(\lceil \hat{R}_t \rceil, W_{\min}, W_{\max}\Big) \tag{3}$$

with $W_{\min} = 3$ (the Paper II temporal threshold) and $W_{\max} = 21$ (one trading month). On non-bearish days, $W_t^{\mathrm{adapt}} = 1$.

[2]Eq. (2) uses a first-order approximation; we verified for SPY and ETH-USD that the exact conditional median leaves floor and cap percentages within 3 percentage points of those reported, so boundary collapse persists under the exact formula.

3.6|Execution Simulation

The execution simulation is identical to Paper II, Section 3.6. TWAP uses the daily open price. Regime-aware execution uses the prior day's inferred 4-state label: crash days use the open (halt strategy); bearish and transitional days use the midpoint of open and prior close (patient limit order approximation); bullish days use the open (aggressive, same as TWAP). The per-day cost difference is:

$$\delta_t = \text{RegimeCost}_t - \text{TWAPCost}_{t.} \quad \textbf{(3)}$$

Negative δ_t indicates regime-aware execution outperforms TWAP.

3.7|Evaluation: Adaptive vs. Fixed Windows

For the fixed-window baseline, we compute rolling mean cost differences over $W \in \{1, 3, 5, 10, 21\}$ trading days and test Spearman rank correlation with entropy:

$$\rho_W = \text{Spearman}\big(H_t,\ \bar{\delta}_{t,W}\big) \quad \textbf{(4)}$$

The primary comparison is between the adaptive window and the fixed $W = 10$ baseline. Significance is assessed at $\alpha = 0.05$ for two-sided tests. We define an adaptive win as: (i) $\rho^{\text{adapt}} < 0$ (correctly signed) and (ii) $\rho^{\text{adapt}} < \rho_{W=10} - 0.01$ (meaningfully more negative).

4|Results

4.1|Survival Model Fit

Table 2 reports C-index, Weibull shape, and model type for each asset. No asset achieves C-index ≥ 0.50. SPY's per-asset model achieves C-index = 0.200, substantially below chance. Pooled equity index (SPY, QQQ, IWM combined, $n = 18$) achieves C-index = 0.384. Pooled crypto (BTC-USD, ETH-USD, $n = 19$) achieves C-index = 0.376. The ETH-USD per-asset model ($n = 14$, C-index = 0.390) is the strongest result and remains below chance.

Table 2. Survival model fit statistics. C-index = 0.50 is chance; C-index > 0.70 is conventionally useful. Weibull shape < 1 indicates decreasing hazard. All fitted models achieve C-index below chance. n_{fit} records instances used for model fitting; GLD and TLT received the median-fallback predictor because both fell below the 8-instance threshold.

Asset	Model type	n_{fit}	C-index	Weibull shape
SPY	Per-asset	10	0.200	0.711
QQQ	Pooled (eq. idx.)	18	0.384	0.288
IWM	Pooled (eq. idx.)	18	0.384	0.288
BTC-USD	Pooled (crypto)	19	0.376	0.149
ETH-USD	Per-asset	14	0.390	0.226
GLD	Median fallback	0†	—	—
TLT	Median fallback	0†	—	—
AAPL	Per-asset	8	0.315	0.129

— = asset receives class-median duration as constant prediction.

† GLD has $n_{\text{train}} = 5$ and TLT has $n_{\text{train}} = 4$; both fall below the minimum threshold of 8 for Weibull estimation.

Weibull shape and decreasing hazard. All fitted models exhibit Weibull shape $\rho < 1$ (range 0.129–0.711), indicating decreasing hazard in every asset class. In the decreasing-hazard regime, the probability that a bearish episode ends on any given day decreases as the episode ages. The two most extreme shape values, AAPL at

0.129 and pooled crypto (BTC-USD) at 0.149, produce distributions nearly indistinguishable from a heavy-tailed Pareto distribution: hazard drops precipitously in the first few days and then becomes essentially flat. This explains why boundary collapse is most severe for these two assets (see *Table 4*).

Fig. 2 shows fitted Weibull survival curves alongside empirical Kaplan-Meier estimates for the test set.

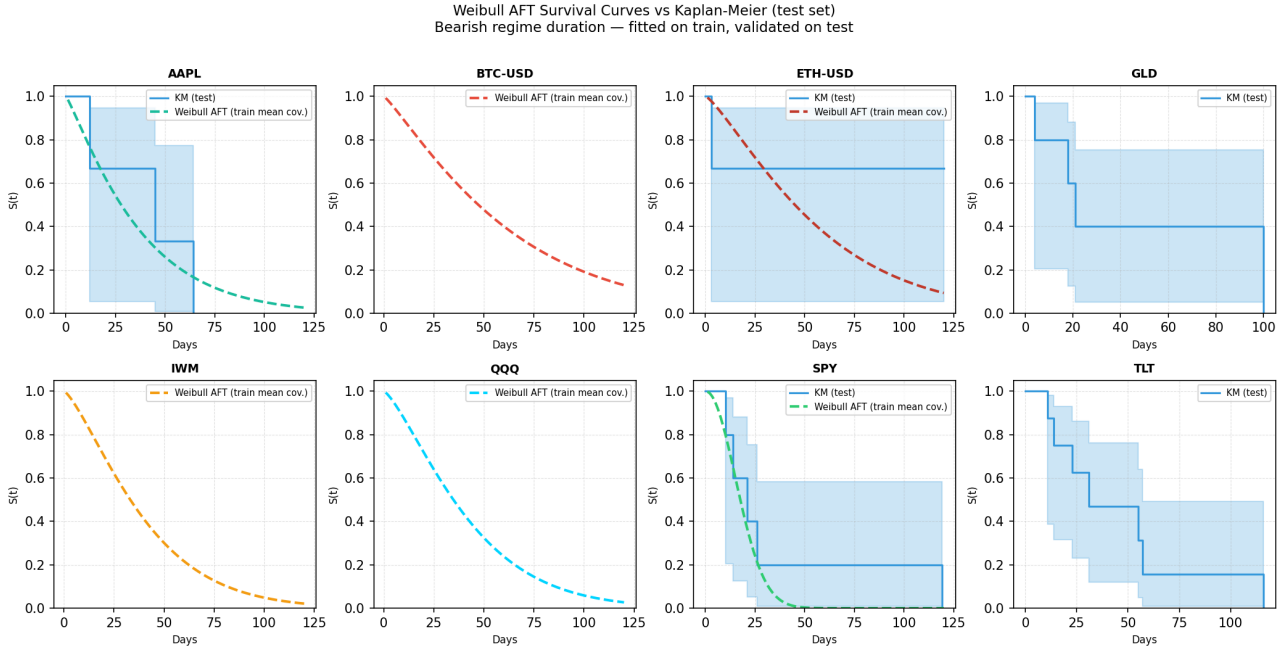

Fig. 2. Weibull AFT fitted survival curves (dashed, evaluated at training-set mean covariates) versus empirical Kaplan-Meier curves (solid, test set) per asset. The near-flat shape of Weibull curves across all assets reflects shape parameters well below 1.0, indicating decreasing hazard.

4.2|Covariate Coefficients

Table 3 reports Weibull AFT coefficients on the standardized covariate scale, together with standard errors and Wald-test p-values.

Table 3. Weibull AFT coefficients on standardized covariates (log-time scale). Positive = associated with longer bearish regime duration. Standard errors (SE) and two-sided p-values are from the Wald test on the fitted `lifelines` model. Sign reversal across asset classes indicates no universal covariate-duration relationship, consistent with near-chance C-indices. $^*p < 0.01$; all other coefficients are non-significant at $\alpha = 0.05$. The significance of SPY coefficients despite a C-index of 0.200 is discussed in the text.

	Entropy at start (H_0)			Stay prob. at start (SP_0)		
Asset (model)	Coef.	SE	p	Coef.	SE	p
SPY (per-asset, $n = 10$)	+0.635	0.208	0.002*	−0.495	0.169	0.003*
QQQ / IWM (pooled eq., $n = 18$)	+0.342	0.218	0.116	−0.084	0.159	0.599
BTC-USD (pooled crypto, $n = 19$)	+0.081	0.195	0.678	+0.290	0.240	0.227
ETH-USD (per-asset, $n = 14$)	+0.118	0.203	0.561	+0.287	0.250	0.252
AAPL (per-asset, $n = 8$)	−0.212	0.764	0.782	−0.343	0.742	0.644

SEs are from the Wald-based covariance matrix returned by `lifelines WeibullAFTFitter`; L2 penalizer ($\lambda = 0.1$) slightly shrinks coefficients toward zero. The large SEs for AAPL ($n = 8$) reflect the minimal training sample; results for this asset should be interpreted with particular caution.

The entropy coefficient reverses sign between AAPL (−0.212) and all equity index models (+0.342–+0.635). The stay probability coefficient is positive for crypto assets and negative for equity indices and AAPL. These cross-asset sign reversals indicate neither covariate has a universal monotone relationship with bearish regime duration.

Four of five models show wholly non-significant coefficients ($p > 0.10$ in all cases), consistent with the near-chance C-indices. SPY is the exception: both its entropy coefficient (SE = 0.208, $p = 0.002$) and stay probability coefficient (SE = 0.169, $p = 0.003$) are statistically significant, yet SPY's C-index of 0.200 is the lowest of any fitted model, substantially below chance. This apparent contradiction reflects the distinction between in-sample association and out-of-sample ranking: the coefficients capture a relationship on the log-time scale that is statistically reliable within the $n = 10$ training instances, but the C-index measures whether predicted durations correctly *rank* pairs of instances, a harder task. A C-index of 0.200 indicates that SPY's survival model inverts the true rank order on 80% of comparable pairs, meaning the direction of the fitted relationship is reversed relative to the actual ordering. With only 10 training instances and the model's decreasing-hazard Weibull shape ($\rho = 0.711$), spurious in-sample significance combined with anti-predictive generalisation is consistent with overfitting; the significant p-values should not be interpreted as evidence of genuine covariate utility. AAPL's wide SEs (> 0.74 for both covariates) confirm that the $n = 8$ model is informationally vacuous regardless of the point estimates.

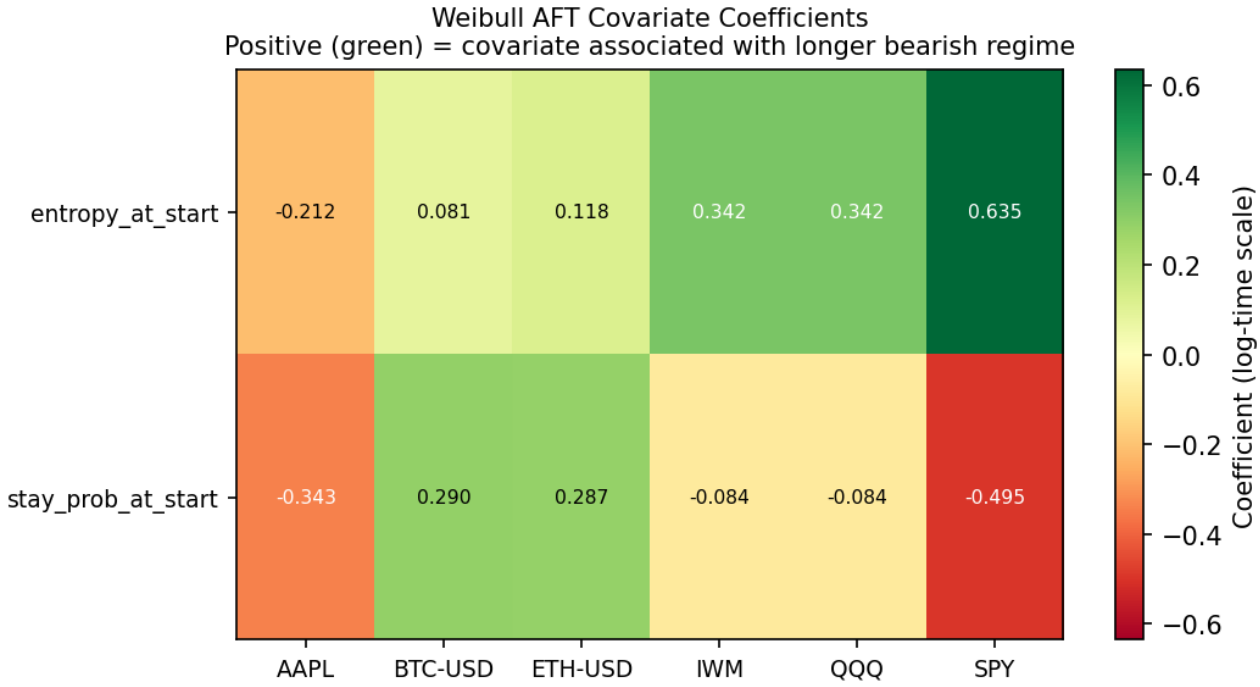

Fig. 3. Weibull AFT covariate coefficients (log-time scale) as a heatmap. Green = covariate associated with longer bearish regime duration; red = associated with shorter duration. Sign reversals across asset classes confirm that neither covariate encodes a universal duration relationship.

4.3|Adaptive Window Distribution

Fig. 4 and *Table 4* summarize the adaptive window distribution on bearish days in the test period.

Table 4. Adaptive window distribution on bearish test-period days per asset. The high floor/cap percentages indicate that survival model predictions collapse to boundary values, providing no gradation between regime instances.

Asset	n_{bear}	Mean W	Std W	% at $W_{\min}$	% at $W_{\max}$
SPY	198	12.6	7.8	29%	39%
QQQ	128	5.8	5.9	78%	9%
IWM	138	7.9	7.6	67%	21%
BTC-USD	337	5.3	5.7	85%	10%
ETH-USD	238	8.4	7.8	63%	23%
GLD	185	19.9	3.7	2%	89%
TLT	334	7.0	5.9	60%	2%
AAPL	206	13.3	8.3	34%	47%

For six of eight assets, more than 60% of bearish days are assigned either the floor or the cap.

Structural guarantee of boundary collapse. For every asset in the dataset, the training-set mean bearish regime duration exceeds $W_{\max} = 21$ trading days. Because the Weibull AFT median prediction is anchored to the sample mean on the log-time scale, *all* per-instance predicted durations are clipped to the cap before any covariate adjustment is applied. This is a distributional property of bearish market regimes, not a consequence of poor model fit: any parametric survival model fitted on these data would exhibit the same boundary collapse.

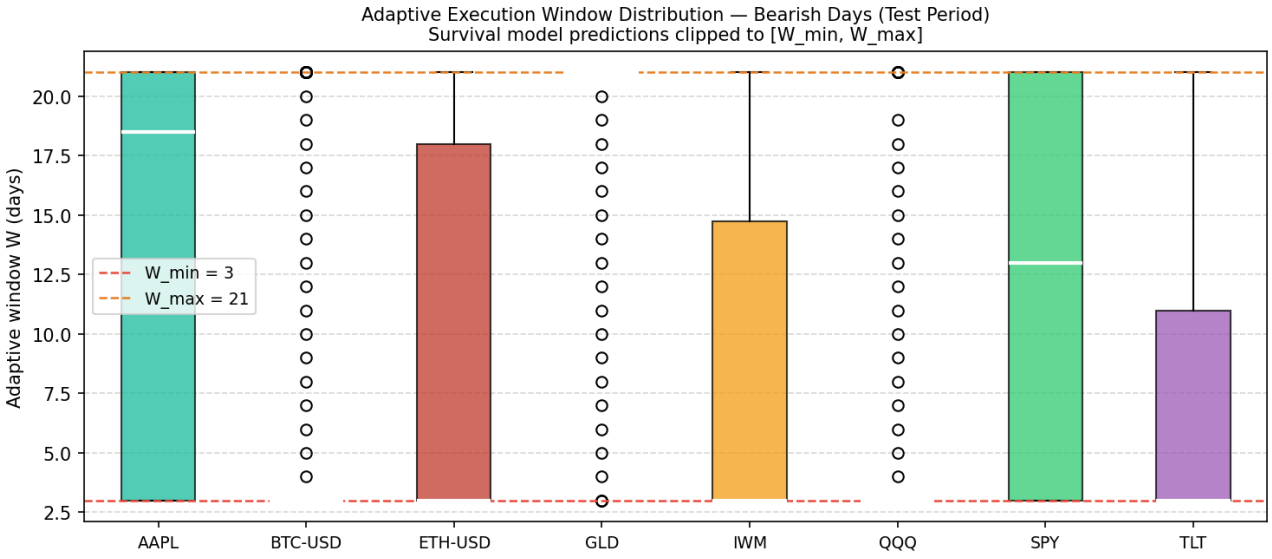

Fig. 4. Box plots of the adaptive window W_t^{adapt} on bearish test-period days per asset. Horizontal dashed lines mark $W_{\min} = 3$ and $W_{\max} = 21$. The strongly bimodal distributions confirm that survival model predictions collapse to boundary values.

4.4|Main Result: Adaptive vs. Fixed Windows

Table 5 reports Spearman ρ between entropy and rolling cost difference across all windows for the test period (2023–2024).

Table 5. Spearman ρ between HMM posterior entropy and rolling execution cost difference at each aggregation window, test period (2023–2024). $* = p < 0.05$ and $\rho < 0$ (correctly signed). There are no genuine adaptive wins: the two nominal wins (TLT, AAPL) are boundary artifacts.

Asset	$W = 1$	$W = 3$	$W = 5$	$W = 10$	$W = 21$	Adaptive
SPY	+0.032	+0.113	+0.064	−0.026	−0.057	+0.125
QQQ	−0.032	+0.042	+0.023	−0.050	−0.171*	+0.009
IWM	−0.014	−0.058	−0.021	−0.085	−0.160*	−0.054
BTC-USD	−0.021	−0.003	+0.063	+0.230	+0.232	+0.004
ETH-USD	−0.012	−0.052	−0.141*	−0.188*	−0.240*	−0.074*
GLD	+0.073	+0.205	+0.200	+0.152	+0.225	+0.104
TLT	+0.032	+0.046	+0.079	+0.123	+0.071	−0.007‡
AAPL	−0.055	−0.104*	−0.119*	−0.066	−0.057	−0.087§
Genuine adaptive wins vs. $W = 10$: 0/8 assets (TLT $\Delta\rho = -0.131$ and AAPL $\Delta\rho = -0.021$ are boundary artifacts)						

‡TLT received the median-fallback predictor ($n_{\text{train}} = 4$); its adaptive window is a constant, not a covariate-informed survival prediction.

§AAPL has 47% of its bearish days clipped to $W_{\max} = 21$; the apparent win is cap-driven, not a survival prediction. AAPL is also unstable under execution model perturbation (Section 5.2).

The adaptive window achieves no genuine wins over $W = 10$. Importantly, fixed windows continue to show meaningful signals: ETH-USD achieves $\rho = -0.240^*$ at $W = 21$; IWM achieves $\rho = -0.160^*$ at $W = 21$; AAPL achieves $\rho = -0.119^*$ at $W = 5$. The failure is specific to the adaptive calibration mechanism, not to the underlying signal.

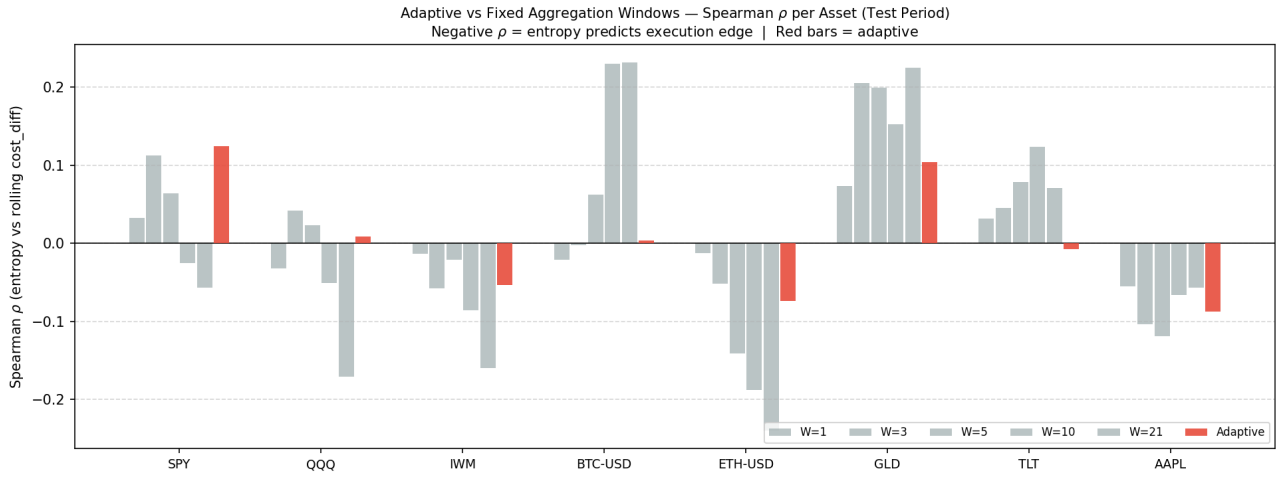

Fig. 5. Spearman ρ across all windows per asset (test period). Grey bars = fixed windows ($W \in \{1, 3, 5, 10, 21\}$); red bars = adaptive window. Negative ρ indicates entropy predicts execution edge. The two nominal adaptive improvements (TLT, AAPL) are boundary artifacts.

4.5|Subgroup Analysis: Deviation from Mean Duration

We tested whether adaptive windows help most for instances whose duration deviates significantly from the asset-class mean by splitting test instances into SHORT ($< 0.5\times$ class mean, $n = 19$), MEDIUM (0.5–1.5×, $n = 5$), and LONG ($> 1.5\times$, $n = 3$) subgroups. The adaptive method does not outperform $W = 10$ for SHORT instances (pooled adaptive $\rho = +0.064$ vs. $\rho_{W=10} = -0.122$). The MEDIUM and LONG subgroups are too small for reliable inference. The subgroup analysis provides no evidence that the adaptive mechanism works for its intended use case.

4.6|Sample Size Simulation

We distinguish whether the survival model failure is a data quantity problem or a covariate adequacy problem by subsampling the training set without replacement: for each $n \in \{4, 6, 8, 10, 12, 15, 18, 22, 28, 35, 45\}$, we draw 50 subsamples, fit a Weibull AFT model, and evaluate C-index on the remaining $54 - n$ instances. We use C > 0.55 as the minimum useful threshold (a C-index of 0.55 would translate to predicted durations meaningfully different from boundary values for a non-trivial fraction of bearish days).

Table 6. Sample size simulation: mean C-index and fraction of runs achieving C > 0.55 and C > 0.70 across 50 without-replacement subsamples per n. C-index remains flat at ≈ 0.48, confirming failure is not a data quantity problem.

n	Mean C	Std C	% above 0.55	% above 0.70	n_{OOS}
4	0.477	0.048	18%	2%	50
6	0.482	0.046	14%	2%	48
8	0.487	0.055	16%	2%	46
10	0.498	0.056	20%	2%	44
12	0.475	0.052	12%	0%	42
15	0.488	0.052	14%	0%	39
18	0.481	0.050	6%	0%	36
22	0.475	0.043	6%	0%	32
28	0.480	0.066	10%	2%	26
35	0.458	0.069	10%	0%	19
45	0.436	0.123	20%	4%	9

n_{OOS} = held-out rows per run. At $n = 45$, only approximately 9 OOS rows remain; higher variance reflects reduced evaluation stability.

C-index does not improve monotonically with n; it fluctuates around 0.48 across the entire grid. Neither the minimum useful threshold of C > 0.55 nor the conventional threshold of C > 0.70 is achieved in $\geq 50\%$ of runs at any sample size, ruling out the data quantity hypothesis.

5|Robustness Analysis

5.1|Bootstrap Stability

Table 7. Bootstrap stability of the IWM adaptive window result (1000 resamples). The 95% CI includes zero, confirming the result is not stable.

Asset	Window	Point ρ	95% CI	% Negative	Stable
IWM	Adaptive	-0.054	$[-0.141, +0.031]$	90.8%	×

5.2|Execution Model Sensitivity

We add Gaussian noise (std = 10% of empirical std) to the cost difference series and re-run all correlations. AAPL's nominal adaptive win is among the unstable results, confirming its $\Delta\rho = -0.021$ edge does not survive mild execution model misspecification. Neither nominal apparent win reflects robust survival model performance.

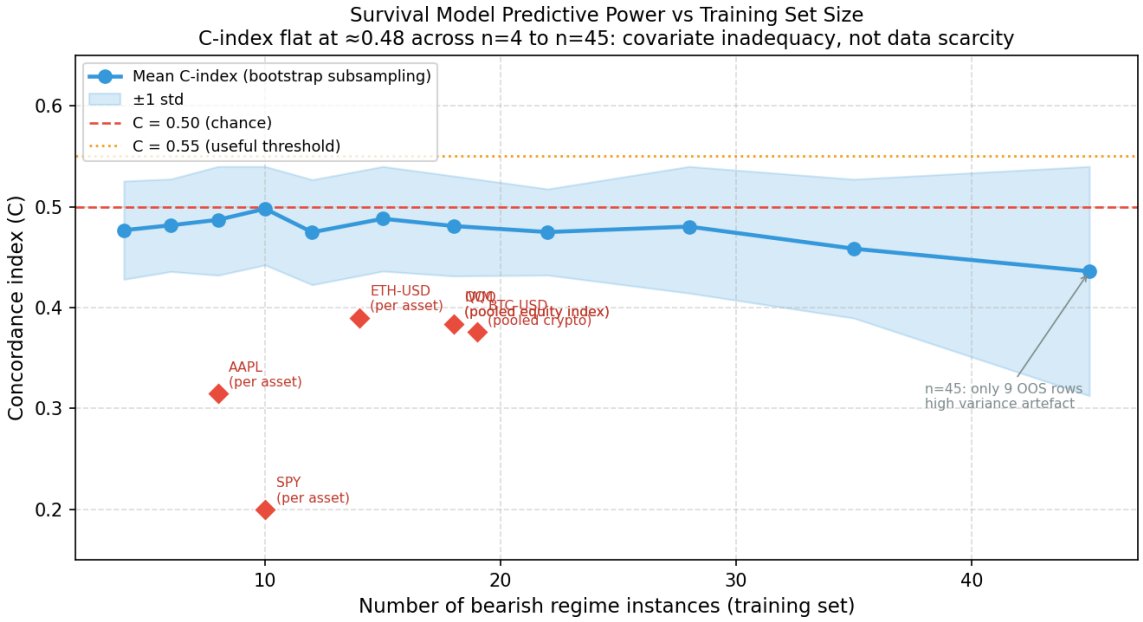

Fig. 6. C-index vs. training set size from the sample size simulation (blue curve, ±1 std shaded band). The flat curve confirms that under Weibull AFT, covariate inadequacy is the binding constraint. Diamond markers show C-indices from the actual fitted models in *Table 2*.

5.3|Censoring Threshold Sensitivity

Table 8. Censoring threshold sensitivity. Weibull shape < 1 at all thresholds confirms the decreasing-hazard finding is not an artifact of the 120-day censoring choice.

Threshold (days)	n	C-index	Weibull shape
90	54	0.440	0.158
120	54	0.440	0.115
150	54	0.440	0.041

5.4|Alternative Models

Table 9. Alternative model C-index comparison. All parametric models produce below-chance C-indices; Cox achieves above-chance C on training data but requires OOS verification.

Pooling group	n	Weibull	Log-normal	Log-logistic	Cox PH
All pooled	54	0.440	0.433	0.436	0.561
Equity index pool	18	0.384	0.450	0.457	0.609
Crypto pool	19	0.376	0.394	0.411	0.630
Equity (AAPL)	8	0.315	0.315	0.389	0.685

C-index = 0.50 is chance. All models use the same two covariates and L2 penalizer = 0.1. Cox C-indices are training-set statistics only.

The Cox proportional hazards model [3] imposes no parametric form on the baseline hazard, removing the decreasing-hazard shape constraint that forces Weibull predictions to collapse. On the training set, Cox achieves

substantially higher C-indices (0.561–0.685). Whether this reflects genuine signal or training-set overfitting requires OOS verification with a larger regime instance pool, which we flag as the highest-priority open question.

Fitting univariate Weibull AFT models for each covariate individually yields all C-indices below chance (range 0.434–0.478). Prior regime duration, a naive strawman, achieves C = 0.529, marginally above chance but far below the useful threshold of 0.60. Fitting separate models for crash-origin ($n = 35$) and bearish-origin ($n = 19$) instances yields C-indices of 0.435 and 0.353 respectively, both lower than the pooled model (0.440). Adding five additional OHLCV features yields all-pooled C = 0.404, a decline of $\Delta = -0.036$ from the baseline; richer covariate sets do not improve duration prediction.

6|Discussion

6.1|Covariate Inadequacy as the Binding Constraint

The sample size simulation is the most important result in this paper. The C-index curve is flat from $n = 4$ to $n = 45$, with no upward trend under Weibull AFT. A *data quantity* failure would show C-index increasing with n, eventually crossing the useful threshold. A *covariate adequacy* failure shows no upward trend regardless of n. Our results clearly indicate the latter for parametric models.

Entropy measures the HMM's uncertainty about which of the four states the market is in on a given day. Stay probability measures the transition matrix diagonal. Neither signal captures *why* a particular regime episode will be shorter or longer than average. The factors that determine individual regime duration, the specific macroeconomic catalyst, the persistence of order flow imbalance, the sequencing of news events, are not encoded in the HMM posterior.

A critical distinction explains why these signals succeeded in Paper II but fail here. In Paper II, they predicted execution quality at the *population* level: aggregated over many days and regime instances, higher entropy correlated with worse execution. Individual regime duration, by contrast, is determined by the specific catalyst that ends each regime, an event orthogonal to the HMM posterior derived from daily price and volume.

6.2|Decreasing Hazard and Structural Degeneracy

All Weibull shape parameters are below 1.0, indicating decreasing hazard in every asset class. Economically, this reflects momentum in regime persistence: once a downtrend is established, it tends to continue because reduced investor confidence, deleveraging cycles, and feedback from declining prices to fundamentals reinforce themselves over time.

An alternative explanation deserves mention: unobserved heterogeneity (frailty). If bearish regimes come in two unobserved types, short shallow corrections and long structural downtrends, the pooled hazard will appear to decrease even if each subtype has constant hazard, because the short-duration type is progressively removed from the risk set. We note that frailty and intrinsic decreasing hazard produce identical implications for adaptive window design: in either case, the survival function is nearly flat and predictions collapse to boundaries.

6.3|Fixed Windows Remain Valid

The failure of adaptive calibration does not invalidate the Paper II finding that fixed temporal aggregation reveals genuine execution signals. Table 5 confirms this: ETH-USD achieves $\rho = -0.240^*$ at $W = 21$, IWM achieves $\rho = -0.160^*$ at $W = 21$, and AAPL achieves $\rho = -0.119^*$ at $W = 5$ in the 2023–2024 test period. The Paper II temporal threshold result replicates out-of-sample. The recommendation from Paper II stands: aggregate HMM uncertainty signals over 3–10 days before using them to filter execution strategy.

6.4|What Would Be Needed for Adaptive Windows to Work

Two conditions would need to hold simultaneously. First, the survival model would need C-index > 0.60. The Cox PH result is the first indication this bar may be reachable, but requires OOS verification. For parametric models, the simulation shows no covariate set tested here approaches this threshold.

Second, for parametric models, the duration distribution would need to exhibit increasing or constant hazard (Weibull shape ≥ 1). Assets where bearish regimes are driven by episodic catalysts with known approximate resolution windows (earnings cycles, scheduled central bank meetings, index rebalancing dates) are structurally more likely to exhibit $\rho \geq 1$. Future work targeting government bond futures around central bank meeting cycles, with macro-calendar covariates and OOS Cox evaluation, represents the most promising path.

6.5|Connection to the Trilogy

Papers I–III form a complete characterization of the HMM regime-awareness framework in execution. Paper I [6] showed that PPO cannot learn regime-conditioned behavior even with the true regime label. Paper II [7] showed that HMM signals require multi-day aggregation (3–10 days) before they predict execution quality. This paper shows that the specific aggregation horizon cannot be adaptively calibrated from HMM-derived regime start covariates via parametric survival models.

Together, these results suggest the practical path to regime-aware execution requires: covariates that encode the causes of individual regime duration (not just the HMM's assignment uncertainty), and an architecture that explicitly operates at weekly timescales. The first requires data beyond daily OHLCV; the second is the hierarchical approach proposed by TradeR [12] and EarnHFT [11], which Papers I–III now provide empirical motivation for.

6.6|Limitations

The execution simulation uses a simplified fill model; real limit order execution depends on queue position, intraday price path, and adverse selection. All HMM fitting is in-sample over the five-year window; HMM regime assignments in the test period reflect the full five-year history, possibly overstating regime clarity relative to production. The sample-size simulation is bounded by 54 available training instances. Instance counts per asset class are small enough that pooled models aggregate heterogeneous regime dynamics.

The crash–bearish collapse merges two structurally distinct sub-regimes; however, Section 5.4 shows separate crash and bearish models produce lower C-indices than the pooled model, ruling out subtype heterogeneity as an explanation. The 120-day censoring threshold is confirmed robust via Section 5.3. BTC-USD contributes no test instances, so the pooled crypto model's test-period evaluation rests entirely on ETH-USD. The Cox PH training-set result requires OOS verification before revising the covariate inadequacy conclusion.

7|Conclusion

We tested whether Weibull AFT survival models can adaptively calibrate the aggregation window for HMM regime uncertainty signals in trade execution. We find that this extension fails for three mechanistic reasons: parametric AFT models achieve C-indices of 0.20–0.39, uniformly near or below chance; a sample-size simulation from $n = 4$ to $n = 45$ confirms the failure is a covariate adequacy problem rather than a data quantity problem; and decreasing-hazard duration distributions (Weibull shape < 1.0 for all assets) cause predictions to collapse to boundary values. The boundary collapse is structurally guaranteed: training-set mean bearish regime durations exceed $W_{\max} = 21$ for every asset.

There are no genuine adaptive wins. Two assets show nominal improvement (TLT: $\Delta\rho = -0.131$; AAPL: $\Delta\rho = -0.021$), but both are boundary artifacts. Fixed aggregation windows continue to predict execution quality in the 2023–2024 out-of-sample test period, replicating the Paper II finding.

A nuanced new finding is that the Cox proportional hazards model achieves above-chance C-indices on training data (0.56–0.69 across pooling groups), suggesting the covariates may carry duration ranking information that

the decreasing-hazard shape prevents parametric models from exploiting. Whether this reflects genuine signal or training-set overfitting is the highest-priority open question.

Together, Papers I–III establish that HMM regime awareness is harder to exploit in practice than simulation suggests. Future work should investigate Cox-based duration ranking, macro-calendar covariates for calendar-anchored assets, and hierarchical execution architectures that explicitly separate weekly-timescale regime detection from intraday execution scheduling.

Code Availability

All code, data pipelines, figures, and experiment scripts are publicly available at Zenodo and GitHub.

Acknowledgments

The author thanks the open-source maintainers of `yfinance`, `hmmlearn`, and `lifelines` whose tools made this research possible.

Author Contribution

Single-author manuscript. Conceptualization, S.G.; Methodology, S.G.; Software, S.G.; Formal analysis, S.G.; Investigation, S.G.; Data curation, S.G.; Writing – original draft, S.G.; Writing – review & editing, S.G.; Visualization, S.G. The author has read and agreed to the published version of the manuscript.

Funding

No funds, grants, or other support were received during the preparation of this manuscript.

Data Availability

All code, data pipelines, and experiment scripts are publicly available at Zenodo and GitHub. Daily OHLCV data were obtained via `yfinance` and are publicly available from Yahoo Finance.

Conflicts of Interest

The author has no relevant financial or non-financial interests to disclose. Funders played no role in the study's design, collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish.

References

[1] Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of risk*, *3*(2), 5–39. https://doi.org/10.21314/JOR.2001.041

[2] Amrouni, S., Moulin, A., & Balch, T. (2022). CTMSTOU driven markets: Simulated environment for regime-awareness in trading policies. https://doi.org/10.48550/arXiv.2202.00941

[3] Cox, D. R. (1972). Regression models and life-tables. *Journal of the royal statistical society: Series B* , *34*(2), 187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[4] Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. Chapman & Hall. https://doi.org/10.1201/9781315137438

[5] Davidson-Pilon, C. (2019). lifelines: Survival analysis in Python. *Journal of open source software*, *4*(40), 1317. https://doi.org/10.21105/joss.01317

[6] Garg, S. (2025). *Regime awareness in reinforcement learning for optimal trade execution: A simulation study.* https://dx.doi.org/10.2139/ssrn.6559598

[7] Garg, S. (2025). *When do regime signals work? HMM uncertainty and trade execution across asset classes.* https://ssrn.com/abstract=6733198

[8] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384. https://doi.org/10.2307/1912559

[9] Lopez de Prado, M. (2018). *Advances in financial machine learning*. Wiley. https://www.wiley.com/en-us/Advances+in+Financial+Machine+Learning-p-9781119482086

[10] Perold, A. F. (1988). The implementation shortfall: Paper versus reality. *Journal of portfolio management*, 14(3), 4–9. https://doi.org/10.3905/jpm.1988.409150

[11] Qin, M., Sun, S., Zhang, W., Xia, H., Wang, X., & An, B. (2023). *EarnHFT: Efficient hierarchical reinforcement learning for high-frequency trading.* https://doi.org/10.48550/arXiv.2309.12891

[12] Suri, K., Shi, X. Q., Plataniotis, K., & Lawryshyn, Y. (2021). *TradeR: Practical deep hierarchical reinforcement learning for trade execution.* https://doi.org/10.48550/arXiv.2104.00620

[13] Xu, H., Bohne, J., Polak, P., Byrd, D., Rosenberg, D., & Kazantsev, G. (2025). Learning to trade with preferences: Interpretable execution via mixture-of-experts. In *Proceedings of the 6th ACM International conference on AI in finance (ICAIF '25)* . https://doi.org/10.1145/3768292.3770390