COVERAGE-INVERSION: CALIBRATION-TRANSPARENT FAIR-VALUE ORACLES FOR CLOSED-MARKET HOURS

A PREPRINT

Adam Noonan[*]
Independent Researcher
adamnoonan@utexas.edu

July 2026

Abstract

A tokenised US equity settles on-chain around the clock; its reference market trades 6.5 hours per weekday. Every night and weekend it emits nothing, yet lending protocols value the collateral anyway: a Friday price held flat, synthetic weekend quotes, or a dispersion number that collapses to zero at the close. No surveyed incumbent surface publishes a verifiable statement of how wrong its served price may be by the next open. We invert the oracle interface: target coverage τ is the consumer-chosen input, the band is the output, and every read carries a receipt any third party can re-derive from public data. The deployed locally-weighted Mondrian split-conformal architecture is evaluated on 5,996 weekend and 22,624 overnight windows across ten symbols (2014–2026). Held out by symbol, coverage at $\tau = 0.95$ is 0.9497 ± 0.0128, all ten passing Kupiec's binomial test of the violation rate; the symbol $\times\ \tau$ grid passes 40/40 Kupiec against GARCH-t's 31/40; and calibration carries to overnight gaps with only the gap selector changed — a property of closed-market hours, not weekends. The residual is disclosed and priced: bands are per-anchor, not full-distribution, calibrated; remaining within-weekend co-breach becomes reserve guidance ($k^* = 3$ at $\tau = 0.95$); reference implementations in Python, Rust, and Anchor.

1 The blind window

Tokenised equities promise something their off-chain originals cannot: composability — collateral in a lending market, an asset in an automated market maker, a leg in an on-chain structured product — settling continuously and without intermediaries. That promise rests on a trust precondition a bearer token cannot inherit from its issuer: that its on-chain price, *at the moment a protocol liquidates, swaps, or values it as collateral*, is reliably pegged to the underlying equity. The capital arriving on the other side of that precondition is measurable on three distinct axes. In *stock*: tokenised real-world assets on-chain sit near $29B of outstanding value [InvestAX, 2026]. In *institutional commitment*: BlackRock's BUIDL alone holds ~$2.5B and is already accepted as on-chain collateral [InvestAX, 2026]. In *flow*: tokenised US equities specifically have cleared tens of billions in cumulative transaction volume — xStocks above $25B [Kraken, 2026], Ondo Global Markets above $7B since its September 2025 launch [CoinDesk, 2026]. What none of this capital can buy today is the precondition itself: a verifiable statement of how far the served price of a tokenised equity may sit from its underlying while the reference market is closed.

The window is structural. A US-listed equity trades 6.5 hours per weekday — 32.5 of the week's 168 hours — and the longest contiguous closed interval, Friday 16:00 ET to Monday 09:30 ET (65.5 hours), recurs every week. Price-relevant information does not stop when the venue does: French [French, 1980] establishes the weekend as a distinct return-generating regime, and Barclay and Hendershott show that after-hours price discovery is non-trivial [Barclay and Hendershott, 2003] and that after-hours effective spreads run 3–4× regular-hours levels [Barclay and Hendershott, 2004] — adverse selection is greatest exactly where a protocol holding tokenised collateral is most exposed. What incumbent oracles supply during this window is *integrity* — a faithful report of the last value, or of a tracking proxy —

[*]ORCID: 0009-0000-7506-6395

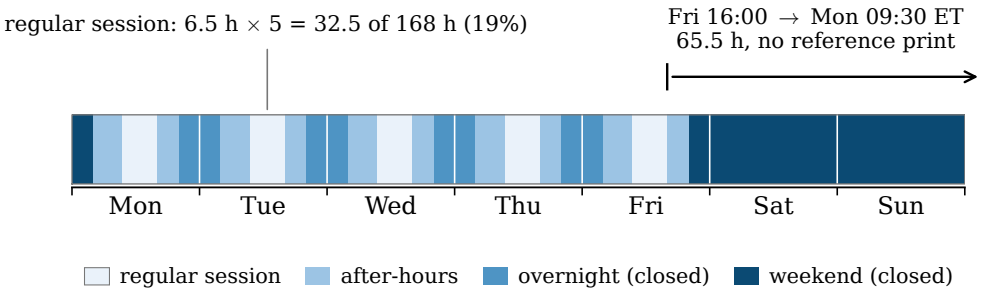

Figure 1: H1a. One wall-clock week (Mon 00:00 → Sun 24:00 ET) for a US-listed equity, coloured by reference-market state, light→dark in order of closedness. The regular session covers 6.5 h × 5 = 32.5 of 168 hours (19%); the Friday-close→Monday-open span leaves 65.5 consecutive hours with no reference print.

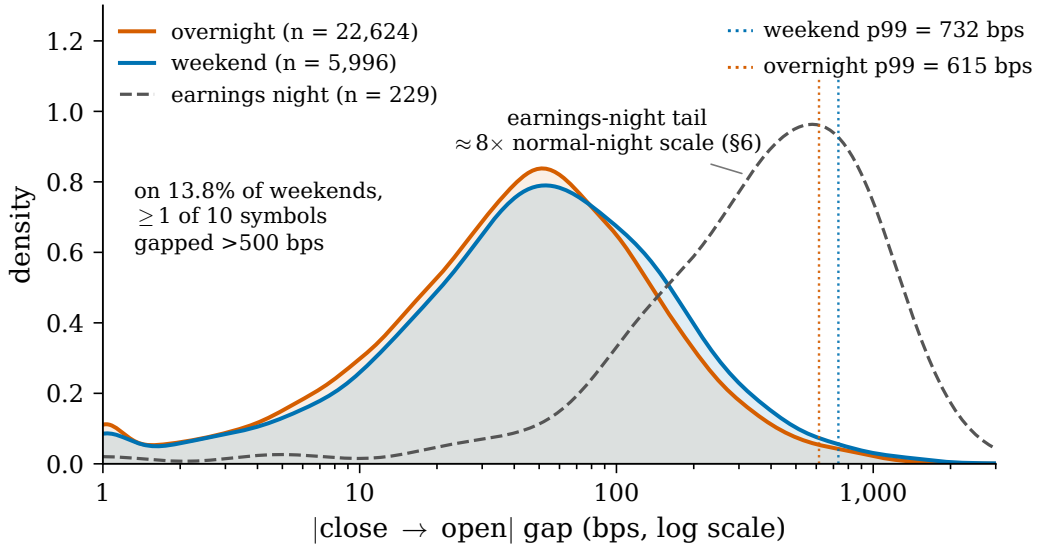

Figure 2: H1b. Realised |close→open| gap distributions (bps, log scale) on the evaluation panel: 5,996 weekend windows (639 weekends × 10 symbols, 2014–2026), 22,624 overnight windows, and the 229 earnings nights among them, whose distribution sits roughly an order of magnitude right of a normal night (the ≈8× of §6 is the calibrated quantile ratio). Dotted lines mark the weekend p99 (732 bps) and overnight p99 (615 bps); on 13.8% of panel weekends at least one of the ten symbols gapped more than 500 bps.

but not a *calibrated statement* of how far that value may sit from fair by the next open. This paper names that gap the blind window, measures what is served into it, and replaces the implicit promise with a contract a consumer can verify.

1.1 The window, measured

How far does the price move while the reference market emits nothing? On the paper's evaluation panel of 5,996 weekend windows (639 weekends × 10 symbols, 2014–2026; §5), the absolute Friday-close→Monday-open gap has a median of 51 bps, a 95th percentile of 319 bps, a 99th percentile of 732 bps, and a maximum of 2,737 bps; 2.2% of symbol-weekends exceed 500 bps and 0.5% exceed 1,000 bps. Aggregating within each weekend across the ten symbols: on 30.8% of the 639 weekends at least one symbol moved more than 300 bps, on 13.8% — about one weekend in seven — more than 500 bps, and on 3.1% more than 1,000 bps. The 22,624-window overnight (close→next-open) panel on the same symbols shows the same shape with thinner tails (median 46 bps, p99 615 bps); weekend tails are uniformly fatter, which is why the paper presents the weekend first and treats overnight calibration as the generalisation test (§6). Figure H1a shows the wall-clock extent of the window; Figure H1b, the realised gap distribution inside it.

Nor is the window a transitional artifact that market-structure reform is about to close. The SEC has granted preliminary approval to the 24X Exchange and to NYSE Arca's 22-hour weekday schedule, and Nasdaq has proposed near-24-hour

weekday trading with clearing infrastructure targeted for late 2026 [Securities and Commission, 2025, Nasdaq, 2025] — but every filed proposal is five-day, retains a daily maintenance pause, and leaves the weekend, the longest and highest-variance closed window measured above, untouched. Extended hours *reshape* the closed-market problem; they do not eliminate it.

The off-hours token process is dual, and both faces are documented on the same panel. Cong et al. [Cong et al., 2025] show that weekend token returns predict the close-to-open underlying return at $\lambda = 0.903$ with adjusted $R^2 = 0.839$ (their Table 10) — a real conditional-mean signal — *and* that extreme moves on token venues "likely would have triggered halts in regulated markets" (their pp. 6–7), because the investor protections of regulated venues are absent regardless of a print's cause. A naive last-trade feed forwards both the signal and the shock; a naive smoother destroys both. The operational consequences are on the record: the JELLYJELLY incident report [HLP, 2025] attributes roughly $12M of Hyperliquid HLP (Hyperliquidity Provider vault) losses (March 2025) to a manipulable single-venue oracle path, and a Volcano Exchange post-mortem [Exchange, 2025] documents $9.21M of Hyperliquid liquidations exceeding the Binance volume that fed the oracle, naming the propagation mechanism in plain language: *uncertainty present in the input was not represented in the output*. Any protocol that accepts tokenised RWAs as collateral must therefore answer, at every block: with the underlying venue closed, what is the fair price — and how uncertain is it?

1.2 What the window serves

No surveyed incumbent surface publishes a calibration receipt verifiable against public data at the aggregate feed level — a statement of the form *"over N historical prediction windows, our 95% band contained the realised price N_{in} times."* What is served into the window instead falls into six archetypes, spanning oracle products *and versions*, summarised in Table T1. Where the wire is decodable we measured the serving behaviour on tape; elsewhere we characterise it from primary documentation and on-chain configuration, and the table flags the evidence class per row so that no measurement is implied that we do not have.

Archetype	Representative	What a consumer sees off-hours	Evidence class
Stale-hold	Chainlink Data Streams v8 via Kamino Scope `Open` feed [Labs, 2025b, Finance, 2026]; Pyth core outside session [Network, 2021]	Last in-session price held forward: the Chainlink v8 Open feed rejects off-hours refreshes `OutsideMarketHours` (Fri 16:00 → Mon 09:30 ET); Pyth's core aggregate is suppressed and its status leaves `Trading` as publishers drop off at the close, preserving the last value	Documentation
Bounded-deviation	Chainlink v8 via Kamino Scope `AllUpdates` feed [Finance, 2026]	Off-hours `midPrice` inside a fixed ±500 bps clamp around Friday's value; no width on the wire	Documentation (mainnet config)
Synthetic-marker	Chainlink Data Streams v11 [Labs, 2026]	Weekend (`marketStatus = 5`) reports carry non-null `mid/bid/ask`	**Measured on tape**
Publisher-dispersion	Pyth aggregate feed [Network, 2021, 2022]	Dispersion σ alongside the price; collapses toward zero at venue close	**Measured on tape**
Executable-overnight	Pyth Pro / Blue Ocean ATS [Network, 2026, 2025]	Real overnight order book, Sun–Thu 20:00–04:00 ET (no Saturday, no most of Sunday); paid enterprise product	Documentation

Archetype	Representative	What a consumer sees off-hours	Evidence class
Continuous tokenised-tracking	Chainlink v10 `tokenizedPrice` [Labs, 2025a]; CF Benchmarks xStocks Indices [Benchmarks, 2026b]; RedStone Live [RedStone, 2026]; SEDA USA500 [Protocol, 2026]	A single 24/7 scalar and no width on the wire; sources vary — crypto-venue order books (Chainlink v10), a BMR-regulated index (CFB), institutional equity vendors for the underlier ticker (RedStone Live, off-chain), or a multi-asset composite (SEDA USA500)	Documentation

Table T1 — What the blind window serves: six off-hours serving archetypes. *Evidence class* distinguishes behaviours measured on decoded wire tape from behaviours characterised from primary documentation and on-chain configuration; the taxonomy is complete but the measurement is partial, and no row implies a measurement we do not have. Kamino's per-reserve choice between the `Open` (stale-hold) and `AllUpdates` (bounded-deviation) serving modes is a governance-mutable parameter we have not decoded per reserve; rows 1–2 describe the two available modes, not a per-reserve assignment. Full comparator methodology in Appendix F.

The findings, per archetype:

Stale-hold — serves Friday's close for the full 65.5-hour window by construction. The gap it does not represent has p95 = 319 bps and p99 = 732 bps on the panel (5,996 weekend windows, 2014–2026).

Bounded-deviation — the clamp width is a fixed parameter. On the panel, at least one of the ten underlyings moved beyond 500 bps on 13.8% of the 639 weekends, and 2.2% of all symbol-weekends did.

Synthetic-marker — the weekend bid/asks are synthetic bookends: `.01`-suffix markers at 100% incidence. SPYx is a pure placeholder (bid 21.01 / ask 715.01 against a last-traded 713.96); QQQx and TSLAx are bid-synthetic (SPYx $n = 6$, QQQx $n = 6$, TSLAx $n = 7$ decoded reports, Feb–Apr 2026); the single NVDAx sample ($n = 1$) classified as real quotes.

Publisher-dispersion — feeds were available near Friday close for 22.1% of attempts pooled (265 of 1,200 symbol × weekend pairs, 2024+, 15:55 ET probe ± 2 h) and 0% for AAPL, GOOGL, HOOD, and NVDA at the same probe times. A consumer reading σ as a Gaussian standard deviation ($k = 1.96$) realises 10.2% coverage on the available subset; the $k = 50$ that realises 95% is fit in hindsight on the evaluation data. That calibration is the consumer's, not Pyth's — consistent with Pyth documenting σ as a publisher-dispersion diagnostic, and Pyth's ≈95% calibration guidance addresses *individual publishers*, a per-publisher self-attestation rather than an aggregate coverage claim.

Executable-overnight — covers weeknight sessions; does not cover the Friday-close→Monday-open window this paper calibrates.

Continuous tokenised-tracking — methodology disclosure ranges from BMR-regulated (CFB) to qualitative (RedStone Live); the feed tracks the wrapped token, whose closed-market wedge against the underlying is measured next.

None of these surfaces emits a calibrated coverage band, so a coverage-vs-sharpness comparison is not well-defined against them; the comparison this paper draws is categorical — the presence or absence of a verifiable per-read calibration statement — and no incumbent makes a coverage claim these findings could contradict.

The archetypes also stratify on a second axis: *which price* they publish. Stale-hold and publisher-dispersion feeds track the underlying ticker and freeze when its venue closes; the synthetic-marker, bounded-deviation, executable-overnight, and tokenised-tracking feeds stay live on the wrapped token or its crypto-venue proxy — a structurally different instrument. Cong et al. [Cong et al., 2025] (their Table 4) measure the wedge for TSLAx, the Backed SPL that Kamino prices: the tokenised price deviates from the last underlying close by more than 1% in 71% of weekday off-hours and 15% of weekend hours, and by more than 5% in 12% of weekday off-hours — beyond the fixed ±500 bps clamp the bounded-deviation mode enforces. The wedge persists by structural design — issuer-controlled redemption windows, KYC/AML constraints on minting, primary-arbitrageur geofencing (their §3.3.3 and §6.1) — rather than as a transitional inefficiency.

The strongest empirical claim a tokenised-tracking feed could credibly make is the one Cong et al. establish: a 1% off-hour token return predicts a 0.903% close-to-open underlying return. That is a conditional-*mean* statement, not a conditional-*quantile* statement, and it is regime-dependent — informative on weekends, reversal-driven on weekday overnights (their Table 8: highest-minus-lowest-quintile next-day spread of -0.99%, $t \approx -2.7$). The tokenised-tracking archetypes therefore publish an implicit conditional-mean point with no width and no signal of which regime the consumer is in; the underlying-tracking archetypes publish a stale print that ignores the off-hours signal entirely. Neither gives a risk-critical consumer the conditional-quantile statement a consumer-facing coverage SLA would need.

1.3 The adoption asymmetry

The market trades these assets far more than it collateralises them. xStocks has cleared more than $25B in total transaction volume against roughly $225M of aggregate AUM — a two-orders-of-magnitude gap between trading use and holding use [Kraken, 2026]. Where the assets are accepted as collateral, the terms are conservative: Kamino's eight xStock reserves (2026-04-26 on-chain reserve-config snapshot, decoded from the klend program; Appendix F) cap loan-to-value between 73% (SPYx) and 30% (MSTRx, HOODx), and four of the eight reserves have borrowing disabled outright. On the 24/7 perpetual-futures venues that nominally price these tokens, median weekend trade density on xStock-backed perpetuals is 0.10% of one-minute bars (≈4 traded bars per 3,931-minute weekend; maximum 3.1%) on a 118-cell (symbol × weekend) Kraken Futures panel spanning nine symbols and 10–19 weekends per symbol, with several names recording zero weekend trades on the majority of observed weekends (AAPL and GOOGL 6 of 10, GLD 9 of 19). And across 102 liquidations of tokenised-equity collateral on a $4B-TVL Solana lending market over nine months (Aug 2025–Apr 2026), none occurred over a weekend or overnight; 101 of 102 fell in US regular trading hours. These observations are consistent with participants declining to transact or lever these instruments off-hours with confidence — an adoption asymmetry between a >$25B trading market and the collateral use the same assets attract. The asymmetry also renders the common weekend fallback circular: the standard recommendation is to reference a secondary venue's price, but those venues are empirically dormant precisely when their price would be needed.

1.4 Inverting the oracle interface

Our answer inverts the conventional oracle interface. Instead of publishing a point and a band whose coverage is implicit and unauditable, the oracle takes target coverage τ as the consumer's input, serves the band that has *empirically* delivered coverage τ on a rolling historical sample, and attaches a receipt to every read — the tuple $(\tau,\, c(\tau),\, q_r(\tau),\, \hat{\sigma}_s,\, r)$ formalised in §3 — that any third party can re-derive from public data and check against the served band. We call this **coverage inversion**; the rest of the paper is the evidence that it can be deployed, audited, and held to its claim. The scope is closed-market hours generally, not weekends: §6 shows the identical architecture calibrates on overnight (close→next-open) gaps with only its gap selector changed. The overnight cadence also surfaces the sharpest — and uniquely *schedulable* — instance of the problem: the earnings night. Because an earnings release is a publicly dated event, the band can widen for it deterministically and in advance — **calendar-conditioned coverage** (§4, §6) — and stays calibrated against an earnings-night tail roughly 8× the normal-night scale.

The binding consumer set is narrower than "any protocol holding tokenised RWA." Three categories carry structural off-hours underlying-side exposure. Lending protocols determine solvency from the future redemption value of collateral rather than its current crypto-venue quote — Kamino's TSLAx-as-collateral integration on a $4B-TVL market is the live anchor [Cong et al., 2025]. Options and structured-product issuers need an explicit forward distribution on the underlying to price strikes and hedge; the conditional-mean passthrough is necessary but not sufficient. Regulator-facing risk models under SR 11-7 / SR 26-2 [of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011, Board of Governors of the Federal Reserve System, 2026] need a price-feed methodology defensible to a model-risk officer. For perp settlement marked to tokenised spot (Kraken xStock perpetuals via CFB's regulated xStocks Indices [Benchmarks, 2026b]) and for DEX market-making, the existing point feeds are correct by contract design, and the calibration primitive is over-engineering.

1.5 Contributions

The paper makes four claims.

- **K1 — the taxonomy (§1.2, §2; Appendix F).** The first tape-backed taxonomy of off-hours oracle serving behaviour: six archetypes spanning oracle products and versions, each row flagged as measured on tape or characterised from documentation. Where the wire is decodable, we measure: synthetic weekend bookends on Chainlink v11 (`marketStatus = 5`, Feb–Apr 2026), Pyth feed availability of 22.1% pooled near Friday close (265/1,200 attempts, 2024+), and a full multiplier ladder showing that no ex-ante-choosable scaling of published dispersion yields a known coverage level.

- **K2 — the interface (§3).** Coverage inversion as an oracle primitive: target coverage as a published, parametric product input, with a per-read receipt re-derivable by any third party from public data. Calibration becomes auditable and SLA-shaped rather than promised.
- **K3 — the architecture and its evidence (§4–§7).** A simple, fully-disclosed locally-weighted Mondrian split-conformal architecture delivers the contract on a 12-year public-data panel: leave-one-symbol-out coverage at $\tau = 0.95$ of 0.9497 ± 0.0128 with all ten held-out symbols passing Kupiec — a binomial test of the violation rate — a nested temporal holdout realising 0.9504, and 40/40 Kupiec on the symbol × τ grid against GARCH-t's 31/40. The same architecture calibrates overnight with one selector changed, and widens *ahead* of publicly dated earnings while staying calibrated. §7 isolates the three load-bearing components.
- **K4 — the disclosed failure mode (§8).** Pooled DQ — a regression test that violations are unpredictable — rejects: the bands are *per-anchor*, not full-distribution, calibrated. The residual is within-weekend cross-sectional co-breach, quantified into reserve guidance ($k^* = 3$ at $\tau = 0.95$). Publishing that measurement requires a calibration-transparent oracle; none of the surveyed surfaces publishes the inputs from which a consumer could compute it.

We do not claim point-accuracy or absolute bandwidth dominance over any incumbent surface; the contribution is the receipt and its per-symbol generalisation. §2 positions the work against conformal prediction, VaR backtesting, and oracle design; §3 formalises the contract; §4 gives the architecture; §5 the data and regimes; §6 the evidence that the contract holds; §7 what is load-bearing; §8 where it fails; §9 concludes with extensions.

2 Related Work

Three literatures each hold one piece of the contract this paper deploys, and the intersection of the three is empty. Conformal prediction supplies coverage machinery with finite-sample guarantees; it reaches an on-chain wire in one prior construction we are aware of, as Byzantine-robust consensus across live reporting sources [Park et al., 2023] — an integrity mechanism, not a coverage claim. VaR backtesting supplies the audit vocabulary — every test §6 runs — but presumes a stated coverage claim to test, and no deployed oracle makes one. Oracle design supplies the delivery infrastructure, organised entirely around integrity rather than calibration. No prior work, academic or deployed, lets a consumer choose a coverage level over a closed-market window and verify they received it.

Conformal prediction and calibrated forecasting. The organising objective is Gneiting, Balabdaoui, and Raftery's: maximise sharpness subject to calibration [Gneiting et al., 2007] — the criterion §7's ablation applies. The deployed architecture is a locally-weighted Mondrian split-conformal predictor [Vovk et al., 2003, 2005]: per-regime empirical quantiles fit on per-symbol $\hat{\sigma}$-standardised residuals, preserving finite-sample marginal validity within each regime under within-regime exchangeability. Conformalised quantile regression [Romano et al., 2019] is the closest published shape for the continuous conformal upgrade sketched in §9; Barber et al. extend conformal validity to non-exchangeable data via weighted quantiles [Barber et al., 2023], and Xu–Xie's EnbPI targets time series directly [Xu and Xie, 2021]. Allen et al. formalise tail calibration — a forecaster can be calibrated on average yet miscalibrated in the tail [Allen et al., 2025] — precisely the distinction the $\tau = 0.99$ anchor tests. Gibbs, Cherian, and Candès give the framework subsuming marginal, group, and localised conditional coverage [Gibbs et al., 2025]; the deployed taxonomy is its discrete-partition case, chosen because every category must be computable on-chain from published state rather than for statistical optimality. A 2026 thread applies this machinery to one-sided VaR directly and is the nearest methodological neighbourhood to what is deployed here: Schmitt weights the calibration sample by exponential time decay and regime similarity, bounding the coverage gap for arbitrary data-driven weights by conditioning on the regime path instead of assuming weighted exchangeability [Schmitt, 2026]; Zhong makes the exponent on the volatility proxy an explicit design parameter, interpolating between an additive shift and a fully proxy-scaled correction [Zhong, 2026]; others conformalise a quantile-regression-forest VaR [Wang et al., 2026], adopt the same Kupiec/Christoffersen evaluation pair used in §6 [Aich et al., 2025], or construct one-sided intervals carrying tail-specific as well as global validity [Cuonzo and Deliu, 2026]. That thread weights or rescales a single pooled calibration sample where this paper partitions it; it evaluates index- and portfolio-level series, where the cross-sectional heterogeneity driving the per-symbol claim of §6 is aggregated away before calibration; and it benchmarks against the online-conformal family [Gibbs and Candès, 2021, 2022, Zaffran et al., 2022, Angelopoulos et al., 2023] — a comparator set §7 adopts, finding that those methods reproduce the same per-symbol failure as an unweighted pooled quantile unless they are given a per-symbol channel. What this literature lacks is a serving surface: its guarantees have not been packaged as an on-chain interface with a per-read receipt.

VaR backtesting and institutional model risk. Banking supervision solved the evaluation problem this paper inherits. Kupiec [Kupiec, 1995] introduced the unconditional-coverage test — a likelihood-ratio (LR) test of whether the realised rate of band violations matches the claimed rate; Christoffersen [Christoffersen, 1998] decomposed conditional coverage into the Kupiec test plus an independence test on the violation sequence. Engle and Manganelli's dynamic quantile

(DQ) test [Engle and Manganelli, 2004] is the strictest member of the family — a regression test that violations are unpredictable from information available when the band was served — and the one §8's failure analysis turns on. The density-forecast generalisation evaluates the probability integral transform (PIT: the forecast's cumulative distribution evaluated at each realisation, which is i.i.d. uniform when the forecast distribution is correct) [Diebold et al., 1998, Berkowitz, 2001]; our fixed-τ coverage check is a marginal slice of that framework, adopted because the consumer contract is per-τ. The Basel traffic-light framework [on Banking Supervision, 1996] is the Kupiec test in zone-graduated regulatory form, and SR 11-7 / SR 26-2 [of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011, Board of Governors of the Federal Reserve System, 2026] specify the validation, effective-challenge, and ongoing-monitoring bar a bank must clear before consuming a model in a collateral pipeline. Every instrument in this toolkit attaches to a coverage claim; incumbent oracles emit none, so the toolkit has had nothing on-chain to grip. §6 applies it to an oracle read that makes the claim explicitly — and the receipt is designed so a third party can rerun the tests without our cooperation.

Decentralised oracle design. The deployed landscape is organised around an integrity primitive: deliver a faithful value on-chain without a trusted intermediary. Chainlink aggregates node-operator reports [Breidenbach et al., 2021], with Data Streams adding market-status-flagged report schemas for RWAs [Labs, 2024]; Pyth aggregates first-party publisher (price, confidence) submissions [Network, 2021, 2022], extended overnight through the Blue Ocean ATS integration [Network, 2026, 2025]; CF Benchmarks publishes FCA-supervised per-second index scalars for tokenised equities [Benchmarks, 2026b,a]; SEDA serves session-aware feeds under an explicit cost-of-carry continuity rule [Protocol, 2026]; RedStone blends institutional and perpetual-market data into a 24/7 feed [RedStone, 2026]; Switchboard generalises aggregation into permissionless queues [Switchboard, 2023]; UMA and Tellor replace continuous aggregation with dispute-and-stake [Project, 2020, Tellor, 2023]. None of these surfaces publishes a calibration claim on the served value — no schema carries a field asserting the rate at which the truth falls inside anything — so a coverage-versus-sharpness comparison against them is not well-defined and is not reported; per-venue wire-format evidence is catalogued in Appendix F. The closest approach is Pyth's ~95% guidance at the individual-publisher level, which is not a property of the aggregate a protocol actually reads [Network, 2022]. The academic literature concentrates on the same primitive: Angeris and Chitra prove when CFMMs are incentive-compatible price oracles [Angeris and Chitra, 2020], and Eskandari et al.'s SoK catalogues manipulation attacks [Eskandari et al., 2021] — the framing is adversarial (did the attacker corrupt the value?) rather than statistical (over N windows, did the served band contain the truth at the claimed rate?), with the BIS framing decentralisation-versus-efficiency [Duley et al., 2023] and adversarial transaction ordering amplifying the cost of any systematic miscalibration [Daian et al., 2020]. Luo et al.'s systematisation of the tokenised-asset stack keeps the same orientation, treating the reference price as a system input rather than a quantity carrying its own calibration statement [Luo et al., 2026]. Park et al. come closest to joining calibration machinery to the wire: ACon² derives a consensus set across multiple oracle contracts by online conformal prediction, guarantees correctness under distribution shift and Byzantine adversaries, and ships a Solidity implementation [Park et al., 2023]. The output is interval-valued and the machinery is conformal, yet the primitive is still integrity — the width of that set measures how far apart live, observable reporters stand and whether some of them are lying, an uncertainty that shrinks as honest reporters are added. The uncertainty priced here is the opposite kind: irreducible, structural, and owed to a single reference market being closed, where no quorum of honest reporters can observe the reopen. ACon² also serves no consumer-chosen coverage level and emits no re-derivable receipt. Integrity and calibration are orthogonal, and an oracle can be integrity-perfect yet calibration-meaningless: Chainlink faithfully reporting Friday's 16:00 close on Sunday at 14:00 is exactly what it claims to be — and is not asserted to be, and is not, a 95%-coverage estimate of Monday's open. Closest to this paper, Cong et al. document that closed-market tokenised-equity prices carry a conditional-mean signal on the next open — off-hour token returns pass through to the close-to-open underlying return at $\lambda = 0.903$ [Cong et al., 2025]; the wedge magnitudes are tabulated in §1.2 (Table T1). The conditional-quantile complement — a band over the same window whose coverage rate is claimed ex ante and verifiable ex post — appears nowhere: not in their panel, not on any surveyed wire. That complement is this paper.

3 The Primitive

3.1 Setup and notation

Let $\mathcal{S} = \{s_1, \ldots, s_K\}$ be a finite universe of assets (in our evaluation, ten US equities, ETFs, and hybrid crypto-proxies) and let $t \in \mathbb{N}$ index a sequence of *closed-market prediction windows* $[t_{\text{pub}}, t_{\text{tgt}}]$, each the interval between a venue close (t_{pub}) and its next open (t_{tgt}). We evaluate two instances: the **weekend** window (Friday 16:00 ET $\rightarrow$ Monday 09:30 ET), the primary panel of §6, and the **overnight** window (close $\rightarrow$ next-day 09:30 ET open), §6's generalization panel.

For each (s, t) we observe:

- $P_t(s) \in \mathbb{R}_{>0}$: the realised reference price at t_{tgt} (the next open).
- $\mathcal{F}_t(s)$: the conditioning information for the window. The features fixed at t_{pub} are $P_{t^-}(s)$ (the close), the implied-vol indices (VIX, GVZ, MOVE), a calendar flag $\text{earn}_t(s) \in \{0, 1\}$ for a scheduled earnings release inside the window (a release in the coming week on the weekend panel; a release dated inside the close→open gap, by BMO/AMC session, on the overnight panel), and a gap-length flag $\ell_t \in \{0, 1\}$ for calendar weekends ≥ 4 days (weekend panel only). The one input that is *not* fixed at t_{pub} is the contemporaneous factor return (E-mini S&P, gold, 10Y Treasury, BTC futures): it is observable *during* the closed window because Globex futures and BTC trade through it, and is the sole input requiring post-publish computation at serve time (§5.3, §5.4).
- A *regime labeler* $\rho : \mathcal{F}_t(s) \to \mathcal{R}$ defined purely on pre-publish information, with a cadence-specific partition: $\mathcal{R}_{\text{weekend}} = \{\texttt{normal}, \texttt{long\_weekend}, \texttt{high\_vol}\}$ and $\mathcal{R}_{\text{overnight}} = \{\texttt{normal}, \texttt{high\_vol}, \texttt{earnings\_night}\}$ (§5).

A *base forecaster* f emits, at each (s, t, q) with claimed quantile $q \in (0, 1)$, a band $\big[L_t^f(s; q), U_t^f(s; q)\big]$.

3.2 The classical oracle problem

A conventional point-plus-band oracle publishes, at each t_{pub}, a triple $(\hat{P}_t, L_t, U_t)$ whose coverage level is *implicit* or *undisclosed*, so the consumer holds no contract on the realised rate at which $P_t \in [L_t, U_t]$. The §1.2 taxonomy (Table T1) confirms this concretely: whether a band is derived from publisher dispersion, a deviation clamp, or a staleness heuristic, the mapping to a realised probability-of-coverage statement is neither asserted nor auditable.

3.3 The coverage-inversion primitive

We define a stricter contract. Fix a base forecaster f, a regime labeler ρ, a discrete claimed-quantile grid $\mathcal{Q} \subset (0, 1)$, and a rolling historical dataset

$$\mathcal{D} = \Big\{\big(P_t(s), \mathcal{F}_t(s), \{(L_t^f(s; q), U_t^f(s; q))\}_{q \in \mathcal{Q}}, \rho(\mathcal{F}_t(s))\big) : t \in \mathcal{T}_{\text{hist}}\Big\}.$$

The **empirical calibration surface** is the mapping

$$S^f(s, r, q) \;=\; \frac{1}{|\mathcal{T}_{s,r}|} \sum_{t \in \mathcal{T}_{s,r}} \mathbf{1}\big[\, L_t^f(s; q) \leq P_t(s) \leq U_t^f(s; q) \,\big],$$

where $\mathcal{T}_{s,r} = \{t \in \mathcal{T}_{\text{hist}} : \text{symbol} = s,\ \rho(\mathcal{F}_t(s)) = r\}$.

Given a *consumer-chosen* target coverage $\tau \in (0, 1)$, the oracle serves the band at

$$q_{\text{served}}(s, r, \tau) \;=\; (S^f)^{-1}(s, r, \tau),$$

with $(S^f)^{-1}$ defined by linear interpolation between bracketing grid points and a documented fallback to a pooled surface $S^f(\cdot, r, q)$ when $|\mathcal{T}_{s,r}| < n_{\min}$. The deployed grid anchors τ at four audited targets, $\{0.68, 0.85, 0.95, 0.99\}$; an off-anchor request is served by interpolation and disclosed as interpolated rather than separately validated (§8).

The oracle read is the tuple

$$\texttt{PricePoint}(s, t, \tau) \;=\; \big(\hat{P},\ L,\ U;\ \tau,\ c(\tau),\ q_r(\tau),\ \hat{\sigma}_s(t),\ r,\ f\big),$$

the band plus a six-field calibration **receipt**. Authoritatively, the receipt fields and their wire names are: the target coverage τ (`target_coverage`, echoed as `claimed_coverage_served`, equal to the request since the deployed shift $\delta \equiv 0$); the calibration buffer $c(\tau)$ (`diagnostics.c_bump`); the per-regime conformal quantile $q_r(\tau)$ (`diagnostics.q_regime_lwc`); the pre-window per-symbol scale $\hat{\sigma}_s(t)$ (`diagnostics.sigma_hat_sym_pre_fri`); the regime label r (`regime`); and the base-forecaster identifier f (`forecaster_used`). **Two verification contracts, and a receipt must say which is in force.** Everything above describes the deployed contract: the band is a deterministic function of a *pre-committed* artefact, so a consumer downloads one SHA-256-stamped file and re-derives any read in one step, and the provider cannot revise what it published. A characterised extension for the overnight profile lets the per-cell level adapt to realised outcomes (Appendix G), which changes the signature — the band becomes a function of the frozen artefact *and* a state m_r that is itself a deterministic

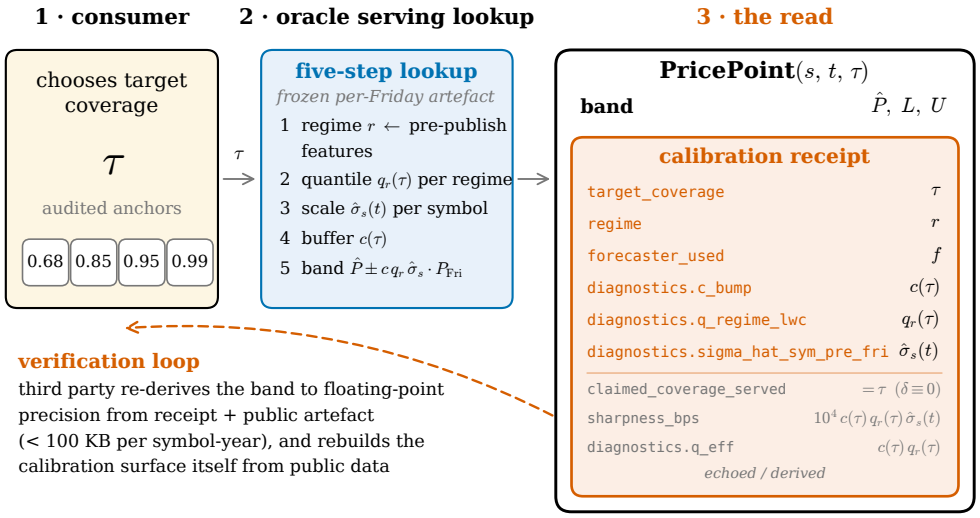

Figure 3: H2 — Anatomy of a read. A consumer chooses the target coverage τ at one of four audited anchors $\{0.68, 0.85, 0.95, 0.99\}$; the oracle resolves the five-step lookup against the frozen per-Friday artefact and returns the band $(\hat{P}, L, U)$ together with the calibration receipt — τ (`target_coverage`), r (`regime`), f (`forecaster_used`), and the diagnostic triple $c(\tau)$, $q_r(\tau)$, $\hat{\sigma}_s(t)$ (`diagnostics.c_bump`, `diagnostics.q_regime_lwc`, `diagnostics.sigma_hat_sym_pre_fri`). Because the receipt fields are sufficient statistics for the band, a third party re-derives L and U to floating-point precision from the receipt plus the public artefact (under 100 KB per symbol-year) and can rebuild the calibration surface itself from public data: coverage is the input, and the receipt makes the claim checkable.

function of public prices. Calibration transparency survives: m is one scalar per regime cell under a published update rule, so anyone can recompute the whole trajectory. What does not survive unaided is the *pre-commitment*, because the provider recomputes m each period rather than having published it in advance. Periodically publishing a hashed checkpoint of m restores it: a read verifies in one step against the checkpoint, auditing the state costs a replay bounded by the checkpoint interval rather than by time since inception, and a provider that deviates from the published rule is caught by recomputing the checkpoint from public prices. The receipt therefore carries the contract type, and under the adaptive profile the multiplier and checkpoint hash are receipt fields rather than metadata — a receipt that omitted them would be insufficient to reconstruct the band, which is the one thing a receipt must never be. Only the overnight profile is a candidate for this; the weekend results in §6 and §7 are served entirely under the pre-committed contract.

The contract is **direction-agnostic**: nothing above requires the served interval to be symmetric about the point estimate. Symmetry enters only through the absolute-value conformity score of §4.2, and Appendix G instantiates the same contract with a one-sided downside score for lending consumers — quantifying, in the process, that a two-sided band at τ delivers $(1+\tau)/2$ downside protection rather than τ. The half-width is echoed in basis points (`sharpness_bps`) and the composite $q_{\text{eff}} = c(\tau)\, q_r(\tau)$ as a derived convenience field (`diagnostics.q_eff`). Under the deployed architecture (§4), the receipt fields are sufficient statistics for the band — q_{served} realises as the factorisation $c(\tau)\, q_r(\tau)$ applied at the per-symbol scale $\hat{\sigma}_s(t)$, giving

$$L = \hat{P} - c(\tau)\, q_r(\tau)\, \hat{\sigma}_s(t)\, P_{t^-}, \qquad U = \hat{P} + c(\tau)\, q_r(\tau)\, \hat{\sigma}_s(t)\, P_{t^-},$$

so a verifier recomputes the band from the receipt to floating-point precision, and re-derives the surface behind it from public data. Figure H2 traces this round trip — τ in, band plus receipt out, third-party re-derivation of S^f and the served band from public inputs alone.

3.4 Properties we claim

(P1) Auditability. $\mathcal{D}$ is reproducible from public data sources (freely available price feeds, vol indices, and earnings calendars). Given $(\mathcal{D}, f, \rho, \mathcal{Q})$, the surface S^f is deterministic; thus any third party can reconstruct S^f and independently verify that the served band corresponds to the published receipt.

(P2) Conditional empirical coverage. Under the assumption that the conditional distribution $P_t \mid (\mathcal{F}_t,\ \text{regime})$ is stationary across $\mathcal{T}_{\text{hist}}$ and the evaluation set, the realised coverage of the served band satisfies

$$\Pr\big[\, P_t \in [L_t^f(s; q_{\text{served}}),\ U_t^f(s; q_{\text{served}})] \;\big|\; \rho(\mathcal{F}_t) = r,\ s \,\big] \;\longrightarrow\; \tau$$

as $|\mathcal{T}_{s,r}| \to \infty$. Finite-sample deviations are measured, tested (§6), and disclosed as the calibration buffer $c(\tau)$; the power of each test — minimum detectable effect per anchor — is tabulated in Appendix B.

(P3) Per-regime serving efficiency at deployment-tuned parameter budgets. The efficiency criterion is mean bandwidth at matched realised coverage *and* per-symbol Kupiec pass rate. On this criterion, §7's ablation establishes empirical dominance over the deployable alternatives evaluated on this panel — the constant-buffer baseline, the un-standardised Mondrian comparator, and parametric GARCH-Gaussian / GARCH-t baselines — with the band served as the §3.3 factorisation at a sixteen-scalar parameter budget (Appendix A) and the $\hat{\sigma}$ rule itself selected under a pre-registered, multiple-testing-corrected gate (§7; variant ladder in Appendix D).

3.5 Non-goals

We explicitly do not claim:

1. **Optimal point estimates.** $\hat{P}$ is the midpoint of a band calibrated for coverage, not a minimum-variance estimator. We make no claim to improve on incumbent oracles' point accuracy during market hours.
2. **Parametric tail guarantees.** The surface is empirical; shock-regime coverage is bounded above by the fraction of historical observations in the same tail of the forecast distribution (§8).
3. **Distribution-free coverage.** P2 assumes stationarity of the *standardised* conditional distribution; $\hat{\sigma}_s(t)$ provides partial adaptive scaling, but the assumption remains. A conformalised variant (Vovk; Barber et al.) would upgrade the statement to a finite-sample guarantee under exchangeability — an open extension (§9).
4. **Protection against adversarial data feeds.** Coverage transparency is orthogonal to upstream-feed integrity: the primitive ingests upstream price signals and republishes them as a calibrated band, its coverage claim conditional on the integrity of those feeds. It is designed to operate alongside, not in place of, an adversarially-robust price feed.
5. **$\hat{\sigma}$ rule optimality.** The deployed $\hat{\sigma}$ rule (EWMA, half-life 8) is justified by a pre-registered three-gate selection procedure under multiple-testing correction with held-out forward-tape re-validation (§7). We do not claim it is optimal among locally-weighted variants.
6. **Optimal lending-policy defaults.** We do not identify the welfare-optimal liquidation rule, collateral haircut, or target τ for a lending protocol; those require account-weight distributions, protocol-specific loss functions, and outcome semantics outside this paper's evidence.
7. **Universal asset-class scope.** The primitive requires a *continuous off-hours information set* — a public, sub-daily signal correlated with the closed-market underlier, as index futures and VIX supply for equities, gold futures and GVZ for gold, Treasury futures and MOVE for rates, and CDS spreads for credit. Real estate, illiquid commodities, and primarily NAV-priced instruments lack such a signal and are out of scope.

4 Architecture

The deployed architecture is a locally-weighted Mondrian split-conformal band around a factor-adjusted point estimate, and it has four ingredients: a point estimator $\hat{P}_{\text{Mon},r}$, a strictly pre-publish per-symbol scale $\hat{\sigma}_s(t)$ that standardises the conformity score, per-regime conformal quantiles $q_r(\tau)$ fit on the standardised residuals, and a near-identity out-of-sample bump $c(\tau)$. This section presents each ingredient in serving order — the order in which a read resolves them — then the five-step lookup and the receipt every read emits. Figure 4 summarises the pipeline; §7 isolates which components are load-bearing.

4.1 Regime labeler

Every read first resolves a regime label $r = \rho(\mathcal{F}_t) \in \{\texttt{normal}, \texttt{long\_weekend}, \texttt{high\_vol}\}$ via the strict priority cascade defined in §5.5, computed from pre-publish quantities only: the VIX percentile at Friday close and the calendar-gap length. The label selects which quantile column serves the read; nothing downstream of the label is symbol-specific except the scale $\hat{\sigma}_s(t)$.

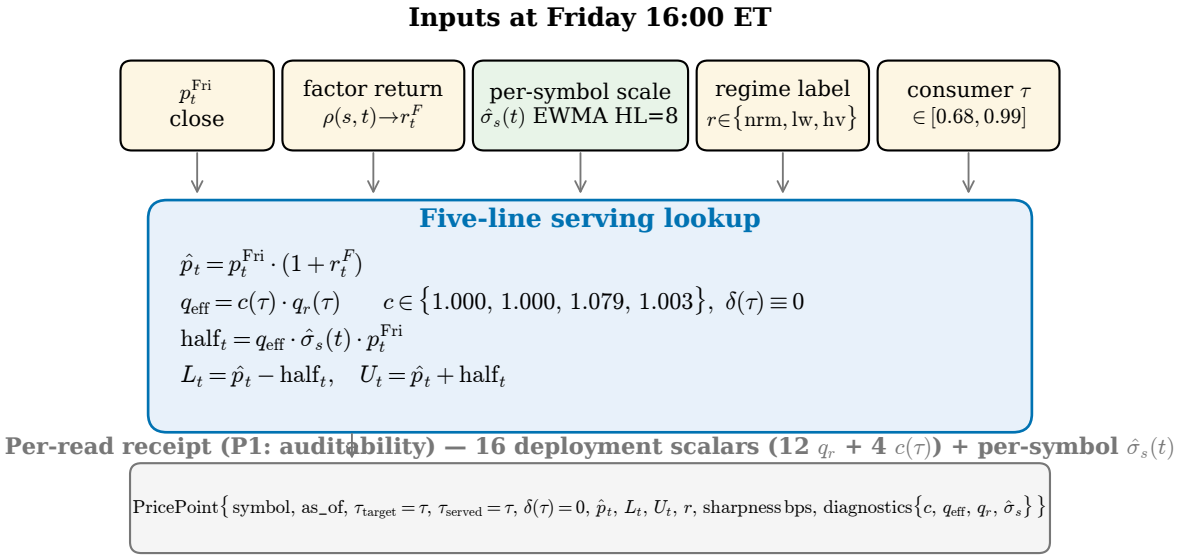

Figure 4: Serving pipeline: five serving inputs — the factor return r_t^F from the §5.4 switchboard (observable *during* the closed window, since Globex futures and BTC trade through it), the regime label r, the per-symbol scale $\hat{\sigma}_s(t)$, the consumer target τ, and the Friday close p_t^{Fri} (the latter four fixed at Friday close) — feed a five-step lookup that returns a band $[L_t, U_t]$ around a factor-adjusted point $\hat{p}_t$. The sixteen deployment scalars are the per-regime quantile table $q_r(\tau)$ plus the bump schedule $c(\tau)$. Every read emits a `PricePoint` receipt carrying the served scalars — the per-read auditability of P_1.

4.2 Point estimator

The point estimator is the per-symbol factor switchboard (§5.4):

$$\hat{P}_{\text{Mon},t}(s) \;=\; P_{\text{Fri},t}(s) \cdot \big(1 + r_t^{\text{factor}}(s)\big),$$

where $r_t^{\text{factor}}(s)$ is the weekend return of the per-symbol factor (ES=F for equities; GC=F for GLD; ZN=F for TLT; BTC for MSTR post-2020-08). The construction encodes a standard result of the price-discovery literature — index futures lead the cash market and carry the dominant information share [Stoll and Whaley, 1990, Hasbrouck, 1995, Gonzalo and Granger, 1995] — and Globex futures and BTC trade through the weekend, so the factor return is observable while the reference market is closed. The point estimator is the input whose residual distribution the conformal quantile is fit on, not the product.

4.3 Per-symbol pre-publish scale $\hat{\sigma}_s(t)$

The conformity score is standardised by a per-symbol estimate of the relative-residual scale, computed strictly from past Fridays:

$$\hat{\sigma}_s(t) \;=\; \text{EWMA}_{\text{HL}=8}\Big(\big\{\, r_{t'}^{\text{rel}}(s) : t' < t \,\big\}\Big), \qquad r_{t'}^{\text{rel}}(s) \;=\; \frac{P_{\text{Mon},t'}(s) - \hat{P}_{\text{Mon},t'}(s)}{P_{\text{Fri},t'}(s)},$$

with half-life $\text{HL} = 8$ weekends (geometric decay $\lambda = 0.5^{1/8} \approx 0.917$ per past Friday). The window is one-sided, strictly before t, so $\hat{\sigma}_s$ is itself a pre-publish quantity. At least eight past observations are required before $\hat{\sigma}_s$ is defined; weekends with fewer are dropped at warm-up (80 of 5,996 panel rows). The half-life was selected under a pre-registered three-gate criterion (Appendix D); constants and the construction algorithm are in Appendix A.

4.4 Mondrian split-conformal on standardised residuals

The conformity score is the relative absolute residual standardised by $\hat{\sigma}_s(t)$:

$$s_t(s) \;=\; \frac{\left|P_{\text{Mon},t}(s) - \hat{P}_{\text{Mon},t}(s)\right|}{P_{\text{Fri},t}(s) \cdot \hat{\sigma}_s(t)}.$$

Weekends partition into the three regimes of §4.1, and the trained per-regime quantile $q_r(\tau)$ is the standard split-conformal $\lceil \tau(n+1) \rceil$-th rank quantile of the regime's training-set standardised scores. The deployed grid has four anchors $\tau \in \{0.68, 0.85, 0.95, 0.99\}$ and three regimes: twelve trained scalars, fit on the 4,186-row pre-2023 calibration set (§5.6), leaving $N \approx 430$–$2,720$ observations per (regime, τ) bin — comfortably above the $N \approx 50$–300 range where Mondrian conformal becomes noisy. With the four $c(\tau)$ bumps below, the complete deployment surface is sixteen scalars (Table A.2).

Standardising the score *before* the Mondrian fit is the choice that makes the per-symbol calibration claim hold. A comparator that skips it absorbs cross-symbol scale heterogeneity — HOOD's typical relative residual is $\sim 3\times$ SPY's — into one cell, producing bimodal per-symbol calibration error: heavy-tail tickers under-cover, low-vol tickers over-cover. Standardisation factors out the per-symbol scale so the regime quantile carries only the *shape* of the residual distribution. The evidence is in §6.4 (10/10 per-symbol Kupiec at $\tau = 0.95$) and the Appendix C simulation study, which reproduces the bimodality under known ground truth and closes it with $\hat{\sigma}$ standardisation on every generating process. Within-bin exchangeability — the assumption yielding finite-sample marginal coverage per cell — is empirically supported (Appendix B).

4.5 The $c(\tau)$ bump

The deployed bump schedule,

$$c(\tau) \;=\; \{0.68\!:\!1.000,\; 0.85\!:\!1.000,\; 0.95\!:\!1.079,\; 0.99\!:\!1.003\},$$

is the smallest $c \in [1, 5]$ per anchor such that pooled out-of-sample realised coverage with effective quantile $c \cdot q_r(\tau)$ matches the target. Three of the four values are essentially identity; only $c(0.95) = 1.079$ carries meaningful out-of-sample information, and §8 carries the provenance disclosure for that one scalar. A companion additive shift $\delta(\tau)$ is identically zero under walk-forward selection (Appendix D) and is retained in the artefact only for receipt-schema compatibility.

4.6 Serving-time band construction

Given a consumer-chosen τ at (s, t), the served band is a five-step lookup:

1. *Read the inputs.* r, $\hat{P}_{\text{Mon},t}(s)$, $\hat{\sigma}_s(t)$, and $P_{\text{Fri},t}(s)$ from the per-Friday artefact.
2. *Look up the quantile.* $q_r(\tau)$, linearly interpolated between the four anchors and clipped outside them.
3. *Apply the bump.* $q_{\text{eff}} = c(\tau) \cdot q_r(\tau)$, with c interpolated on the same anchors.
4. *Construct the band.* $\text{half}_t = q_{\text{eff}} \cdot \hat{\sigma}_s(t) \cdot P_{\text{Fri},t}$; $[L, U] = \hat{P}_{\text{Mon},t} \pm \text{half}_t$. The half-width in basis points is $10^4 \cdot c(\tau) \cdot q_r(\tau) \cdot \hat{\sigma}_s(t)$: heavy-tail tickers receive wider bands than low-vol tickers at identical τ and regime.
5. *Emit the receipt* (§4.8).

Calibration evidence in §6 is reported at the four audited anchors only; off-anchor τ values receive interpolated multipliers with no separate audited evidence, a disclosure §8 carries. There is no surface inversion and no per-symbol fallback — per-symbol behaviour is carried implicitly by $\hat{\sigma}_s(t)$ on the artefact, not by a per-symbol quantile cell.

4.7 Calendar-conditioned coverage

The architecture is parameterised by a gap selector: the weekend panel admits Friday→Monday pairs; the overnight panel of §6.4 admits consecutive-trading-day pairs with a one-line change and no change to any downstream component. The regime partition is re-derived per cadence (definitions and counts in §5.5); the overnight set adds `earnings_night`, the first regime the architecture conditions on *a priori*. Unlike `high_vol` — a market state inferred from trailing VIX — an earnings release is a scheduled event with a publicly known date and session, so the band widens deterministically *ahead* of the information event, before any volatility is realised. We term this **calendar-conditioned coverage**; to our knowledge, no production oracle (Pyth, Chainlink Data Streams, RedStone, Kamino Scope) widens its feed for a scheduled earnings release. Each release is assigned to the single overnight gap it drives via session timing — an after-close report dated t_0 or a before-open report dated t_1 fires inside the close(t_0)→open(t_1) gap. Standardised by the same $\hat{\sigma}_s(t)$ the band uses, the realised move on earnings nights has $p_{99} \approx 9.7$ versus ≈ 2.0 on other nights; the

fitted $q_{\texttt{earnings\_night}}(\tau)$ runs $\approx 8\times$ $q_{\texttt{normal}}(\tau)$, and §6 shows the resulting band is calibrated, not merely wide, with the event-study result in Figure H4.

$\hat{\sigma}$ **de-contamination.** Because an earnings residual can be $\sim 8\sigma$, admitting it to the EWMA scale would over-widen the subsequent half-life of *normal* nights and induce clustered post-earnings violations. Earnings-night residuals are therefore excluded from the $\hat{\sigma}_s$ estimation pool — every night still receives a $\hat{\sigma}_s$ built from its non-earnings history — assigning the earnings excess to the `earnings_night` quantile rather than the scale. Earnings fatness is a regime effect, not a per-symbol scale effect. This separation restores overnight violation independence and collapses the overnight $c(\tau)$ bumps to ≈ 1.0 (§6.4).

4.8 The PricePoint receipt

Every read returns a `PricePoint` carrying the band plus the receipt fields formalised in §3: the target τ, the regime r, the wire-format forecaster tag, sharpness in basis points, and the diagnostic quartet $(c(\tau),\, q_r(\tau),\, q_{\text{eff}},\, \hat{\sigma}_s(t))$. A third party holding the audit-trail sidecar and the per-Friday artefact can re-derive the served band from the receipt alone — read $\hat{P}$, r, $\hat{\sigma}_s$, and P_{Fri} from the artefact, read q_{eff} from the receipt, and verify $L = \hat{P} - q_{\text{eff}} \cdot \hat{\sigma}_s \cdot P_{\text{Fri}}$ and $U = \hat{P} + q_{\text{eff}} \cdot \hat{\sigma}_s \cdot P_{\text{Fri}}$ to floating-point precision. And because every input upstream of the artefact is public (§5.7), the audit does not stop at the band: the same third party can rebuild the calibration surface itself — re-fit the sixteen scalars from public data — and check that the published quantiles are the ones the method produces. Verifying a published band needs only the sidecar and the per-Friday artefact, under 100 KB per symbol-year; the $\hat{\sigma}$ column is recomputable from the same parquet given only the half-life constant, so the artefact is self-contained. This is the operational instantiation of P_1.

4.9 Deployment

The architecture is implemented across three surfaces. The Python `Oracle` (`src/soothsayer/oracle.py`) is the executable specification; a Rust serving stack (`crates/soothsayer-oracle`) reproduces it under a byte-for-byte parity contract — 180/180 verification cases across both forecaster paths (M5, the retired unweighted-Mondrian reference; and M6, the deployed locally-weighted conformal architecture — "LWC" — of §4.1–§4.8; Appendix E) — so a consumer can verify a served `PricePoint` against the published artefact without trusting either implementation. An Anchor program (`programs/soothsayer-oracle-program`) publishes the band on-chain; it is deployed on devnet and currently serves the M5 reference path, with the M6 wire slot reserved and on-chain enablement gated on the next publisher release. The receipt fields that make a band re-derivable are carried in the off-chain `PricePoint` and the public artefact; the on-chain `PriceUpdate` account carries the served band and its coverage bps, and a consumer verifies it against the same public artefact. Wire format, on-chain invariants, and the crate stack are documented in Appendix E.

5 Data

5.1 Universe

We evaluate on $K = 10$ symbols spanning three asset classes: eight US equities (SPY, QQQ, AAPL, GOOGL, NVDA, TSLA, HOOD, MSTR) — the underliers of the Backed Finance xStock tokens on Solana mainnet — plus tokenised gold (GLD) and long rates (TLT), which share the closed-market structure but exercise different factors and volatility indices, so calibration across all three classes evidences generalisation beyond an equity-only universe. The evaluation target is the underlying-venue price — NYSE Monday opens on the weekend panel, next-day opens on the overnight panel — not the on-chain token price: we predict the underlier the token settles against.

5.2 Panels

Each row is one (s, t) close→open window carrying the Friday (or prior-day) close, the next open, the per-symbol factor return, the vol-index close, the calendar-gap length, and an earnings flag (full schema in Appendix A). The **weekend panel** admits pairs of trading days separated by ≥ 3 calendar days: 5,996 rows across 639 weekend dates, 2014-01-17 → 2026-04-24 (HOOD contributes 246 post-IPO weekends), of which 5,916 remain evaluable after the $\hat{\sigma}$ warm-up drop (§4.3). The **overnight panel** applies the same pipeline with the gap selector admitting consecutive-trading-day pairs (§4.7): 22,624 rows across 2,412 weeknights, ≈3.8× the weekend panel. Two cadence-specific treatments apply overnight: each earnings release is assigned to the single gap it drives via BMO/AMC session timing (scryer `yahoo/earnings/v2`), and the 241 ex-dividend-morning opens are reconstructed to their cum-dividend level from `yahoo/corp_actions/v1`, removing the only systematic price-level artefact (Appendix B).

The two panels are different objects, not one architecture applied twice. They differ in gap length (≈65.5 hours against ≈17.5), in event structure (the weekend carries holiday-extended gaps and quarterly expiry; the overnight window carries scheduled earnings, the dominant overnight fat tail, which has no weekend analogue), in regime taxonomy (§5.5 re-derives it per cadence rather than reusing it), and — as §1.4 argues from the filed extended-hours proposals — in expected lifetime. Every filed proposal is five-day and shortens or removes the *overnight* window while leaving the weekend untouched; a 24/5 market has no overnight gap and the same weekend gap. The overnight panel therefore does double duty: it is the generality test for the architecture (§6.8) and simultaneously the surface with the shorter horizon, while the weekend result is the durable one. Reporting both is the hedge, and their divergence under measurement is a finding rather than an inconsistency.

5.3 Pre-publish features

The architecture consumes four features fixed at publish time (Friday 16:00 ET or pre-holiday close): the close $P_{t^-}(s)$, the vol-index close (drives `high_vol`), the long-weekend flag, and the factor return $r_t^{\text{factor}}(s)$. Because Globex futures and BTC trade through the weekend, the factor return is observable at serve time while the reference market is closed — in live deployment it is the *only* input requiring post-publish computation.

5.4 Factor and volatility-index switchboards

The per-symbol mappings are static:

Symbol	Factor	Vol index
SPY, QQQ, AAPL, GOOGL, NVDA, TSLA, HOOD; MSTR pre-2020-08	`ES=F`	`^VIX`
MSTR (2020-08-01 onward)	`BTC-USD`	`^VIX`
GLD	`GC=F`	`^GVZ`
TLT	`ZN=F`	`^MOVE`

The MSTR pivot at 2020-08-01 marks MicroStrategy's first Bitcoin treasury purchase — a structural break that turned a software equity into a leveraged BTC-treasury vehicle, so the factor is reconciled to the asset's actual post-break driver (selection evidence in Appendix D).

The factor return $r_t^{\text{factor}}(s)$ is measured over the closed window itself — from the factor's Friday 16:00 ET close to its contemporaneous print at serve time (the latest Globex/BTC quote when the band is read) — so it spans the window rather than being fixed at publish; it is the only input the point estimator requires post-publish (§5.3).

5.5 The regime labeler ρ

On the weekend panel, ρ is a strict priority cascade: `high_vol` — VIX at Friday close in the top quartile of its trailing 252-trading-day window; else `long_weekend` — `gap_days` ≥ 4; else `normal`. Shares on the 5,996-row panel: normal 3,934 (65.6%), high_vol 1,432 (23.9%), long_weekend 630 (10.5%).

On the overnight panel the partition is re-derived: `long_weekend` has no analog and is dropped; `high_vol` keeps the 252-trading-day VIX lookback; and `earnings_night` is added as the top-priority bucket — a gap is `earnings_night` iff a scheduled release (after-close at t_0 or before-open at t_1) falls inside it. Because the release date and session are public, this is the architecture's only calendar-conditioned regime (§4.7). Shares on the 22,624-row panel: normal 16,604 (73.4%), high_vol 5,791 (25.6%), earnings_night 229 (1.0%).

5.6 Train/test split

Weekends with `fri_ts` before **2023-01-01** are calibration; the remainder is held-out out-of-sample.

Split	Rows	Weekends
Calibration (2014-01 → 2022-12)	4,186	466
Held-out OOS (2023-01 → 2026-04)	1,730	173

The trained per-regime quantiles are fit on the calibration set and held *fixed* throughout evaluation: no post-2023 information enters the quantile table. The $c(\tau)$ bumps are deployment-tuned on the OOS slice — disclosed in §8 — and $\delta(\tau) \equiv 0$ (§4.5). The split places the 2023 banking turbulence and the 2024–2025 macro transition in the held-out slice.

5.7 Provenance

All inputs are read from scryer parquet — `yahoo/equities_daily/v1`, `yahoo/earnings/v2`, `yahoo/corp_actions/v1` — with fetch, retry, and schema ownership in the sibling scryer repository. **Data availability:** all upstream data is publicly available and free (Yahoo Finance daily bars and earnings calendar, CME futures, VIX/GVZ/MOVE, BTC-USD). Soothsayer consumes pre-fetched parquet with `_fetched_at` cutoffs preserved in the metadata; rebuild commands and repository details are in Appendix A.

5.8 Forward tape

The deployed artefact is frozen and content-addressed (SHA-256, freeze date 2026-05-04) and is evaluated on a forward tape appended weekly and never used to re-select any component. As of 2026-07-10 the tape covers 11 post-freeze weekends (110 rows): pooled Kupiec passes at all four anchors and 10/10 symbols pass per-symbol Kupiec at $\tau = 0.95$ (`reports/m6_forward_tape_11weekends.md`). At $N = 11$ the per-symbol tape is not yet powered, however: two symbols (HOOD and MSTR) each realise a 18.2% violation rate (2 of 11) at $\tau = 0.95$, and Kupiec accepts at that n only because the per-symbol test has no power to reject — the per-symbol forward claim rests on accumulation, not on the current pass.

6 The contract holds

6.1 Held out on two orthogonal axes

The $\tau = 0.95$ headline is held out on two orthogonal axes — symbol and time — and passes on both. On the symbol axis, leave-one-symbol-out cross-validation fits each held-out symbol's buffer $c(\tau)$ on the other nine symbols' out-of-sample slice, so the fit never sees the held-out symbol's data: realised coverage is $\mathbf{0.9497 \pm 0.0128}$ across the ten held-out symbols, and all ten pass Kupiec. Heavy-tail tickers (MSTR, HOOD, TSLA) and low-vol tickers (SPY, GLD) alike land in a uniform 0.93–0.97 band when held out — the per-symbol $\hat{\sigma}_s$ carries each held-out symbol's own scale. On the time axis, a nested temporal holdout — TRAIN pre-2023 for the per-regime quantiles, TUNE 2023 for $c(\tau)$, EVAL 2024-01-05 $\rightarrow$ 2026-04-24 untouched by either fit — realises $\mathbf{0.9504}$ (Kupiec $p = 0.947$, per-symbol Kupiec 10/10); $c(0.95)$ fit on TUNE alone matches the full-OOS fit at three decimals, so the headline is not sensitive to fitting on the evaluation slice.

The number is stable in every direction we varied it. Across TUNE anchors $\{2021, 2022, 2023\}$, realised coverage at $\tau = 0.95$ stays in $[0.9474, 0.9524]$; across four OOS-split anchors $\{2021, 2022, 2023, 2024\}$ it stays in $[0.9503, 0.9539]$ with every (split × τ) cell passing both tests; across calendar sub-periods $\{2023, 2024, 2025, 2026\text{-YTD}\}$ the 16-cell grid carries 3 Kupiec rejections — all toward over-coverage, the contract-favourable side — and 0 Christoffersen rejections.

Protocol. The panel is 5,996 weekend windows — ten symbols × 639 weekends, 2014-01-17 $\rightarrow$ 2026-04-24, with the $\hat{\sigma}$ warm-up leaving 5,916 evaluable rows — split into a pre-2023 train slice (4,186 rows) and a 2023+ out-of-sample slice (1,730 rows × 173 weekends) on which everything below is evaluated. Coverage is tested per anchor with Kupiec and with Christoffersen (an independence-of-violations test: breaches should not cluster in time). Pooled OOS at $\tau = 0.95$: realised 0.950 at mean half-width 370.6 bps, Kupiec $p = 0.956$, Christoffersen $p = 0.603$ — both tests pass at every served anchor.

6.2 The master grid

The per-symbol question is whether every ticker's individual cell is calibrated, not just the pool. We evaluate a single 40-cell (symbol × τ) grid under three backtests — Kupiec, Christoffersen, and Engle–Manganelli DQ — against four baselines: a train-fit constant buffer, GARCH(1,1)-Gaussian, GARCH(1,1)-t (the practitioner default), and an unweighted Mondrian comparator (the same per-regime quantile architecture without per-symbol $\hat{\sigma}$ standardisation). Pass counts at $\alpha = 0.05$:

τ	metric	constant buffer	GARCH-N	GARCH-$t^\dagger$	unweighted Mondrian	**this paper**
0.68	Kupiec	4/10	5/10	6/10	1/10	**10/10**
	Christoffersen	7/10	7/10	8/10	9/10	**9/10**
	DQ	2/10	6/10	7/10	1/10	**10/10**
0.85	Kupiec	2/10	8/10	7/10	1/10	**10/10**
	Christoffersen	7/10	7/10	9/10	8/10	**10/10**
	DQ	2/10	8/10	7/10	2/10	**10/10**
0.95	**Kupiec**	**4/10**	**6/10**	**8/10**	**2/10**	**10/10**
	Christoffersen	7/10	7/10	10/10	8/10	**10/10**
	DQ	**4/10**	**6/10**	**8/10**	**5/10**	**10/10**
0.99	Kupiec	7/10	4/10	10/10	8/10	**10/10**
	Christoffersen	5/10	10/10	10/10	4/10	9/10
	DQ	3/10	2/10	9/10	1/10	9/10
40-cell totals	**Kupiec**	**17/40**	**23/40**	**31/40**†	**12/40**	**40/40**
	Christoffersen	**26/40**	**31/40**	**37/40**	**29/40**	**38/40**
	DQ	**11/40**	**22/40**	**31/40**†	**9/40**	**39/40**

† NVDA's t-MLE pushes $\hat{\nu}$ to the optimizer lower bound $\hat{\nu} = 2.50$, adjacent to the $\nu \leq 2$ variance-undefined region; those four cells fall back to GARCH-Gaussian — the practitioner-equivalent of attempting GARCH-t and finding it ill-defined.

The deployed architecture leads every baseline on every test — 40/40 Kupiec, 38/40 Christoffersen, 39/40 DQ — with GARCH-t second on all three (31/40, 37/40, 31/40); at the $\tau = 0.95$ anchor it is the only method on the grid that passes 10/10 on all three tests, with per-symbol violation rates in $[3.5\%, 6.9\%]$ (Fig. S1).

One asymmetry the grid does not itself remove: the deployed method's buffer $c(\tau)$ is fit on the same 2023+ slice the grid scores, whereas the constant-buffer and GARCH baselines carry no out-of-sample coverage correction. The $\tau = 0.95$ column is therefore not an in-sample artefact — §6.1's nested temporal holdout re-fits $c(0.95)$ on 2023 alone and evaluates on the untouched 2024–26 slice, where the per-symbol pass is preserved 10/10 and $c(0.95)$ reproduces the full-slice value to three decimals — but the three remaining anchors are reported at in-sample c, and the tallies should be read as "no per-symbol cell is distinguishable from calibrated at the grid's resolution," with the held-out guarantee established at $\tau = 0.95$ (§6.1) and per-symbol power quantified in Appendix B. The 40-cell / 120-evaluation grid carries no multiple-testing correction, so 40/40 Kupiec is over-conservative-consistent rather than neutral — perfect calibration at $\alpha = 0.05$ would expect ≈ 2 rejections across 40 cells — and the $\tau = 0.99$ per-symbol cells are near-powerless (expected 1.73 violations/symbol), inflating the pass tallies; Appendix B carries the per-cell power.

The GARCH-t contrast is instructive because it is the strongest of the four. Pooled at $\tau = 0.95$, GARCH-t realises 0.928 against its claimed 0.95 (Kupiec $p < 0.001$); the deployed band realises 0.950 ($p = 0.956$). Student-t innovations repair the $\tau = 0.99$ tail, but at the fitted $\hat{\nu} \approx 3$ the standardised-t 95th percentile (~ 1.83) sits *below* the Gaussian 1.96, so the bands tighten exactly where Gaussian was already under-covering — t-innovations fix the wrong end of the distribution. One concession the comparison owes the reader: widened post hoc to matched 95% realised coverage, GARCH-t's half-width is ≈ 378 bps, comparable to the deployed 370.6 bps. So at $\tau = 0.95$ the deployed advantage is not width — it is that the deployed band's coverage is stated in advance and holds ($p = 0.956$), while GARCH-t's ≈ 378 bps is the width that *would have* matched coverage, recoverable only after observing the outcomes it was supposed to bound. At $\tau = 0.99$ the deployed band dominates on coverage and matched-coverage width simultaneously.

6.3 The tokenized-tracking head-to-head

The remaining objection is "why not just track the 24/7 token price?" We construct a proxy for the continuous-tracking archetype — a constructed baseline, not any vendor's product — following Cong et al.'s published passthrough ($\lambda = 0.903$): the live tokenized-perp price as the point, an empirical-quantile residual band around it, evaluated at its most-informed closed-window snapshot. As deployed, at matched coverage (0.949 vs 0.943, both passing Kupiec on $n = 117$ weekend-symbol observations), the deployed band is **45% narrower** in half-width and scores **46% lower** on Winkler (a proper scoring rule for interval forecasts, charging width plus a penalty for misses).

That deployment figure, however, rests on the deployed band's 12-year calibration record against a baseline the live perp tape can only calibrate over a few months — so it conflates architecture with calibration-history length. To isolate the architecture, we re-run both methods walk-forward on the *same* post-2025-12 window (Appendix D): the width edge falls to $\approx$ 27–30% and the Winkler edge to $\approx 20\%$ with a bootstrap 95% CI that **includes zero** (matched-history

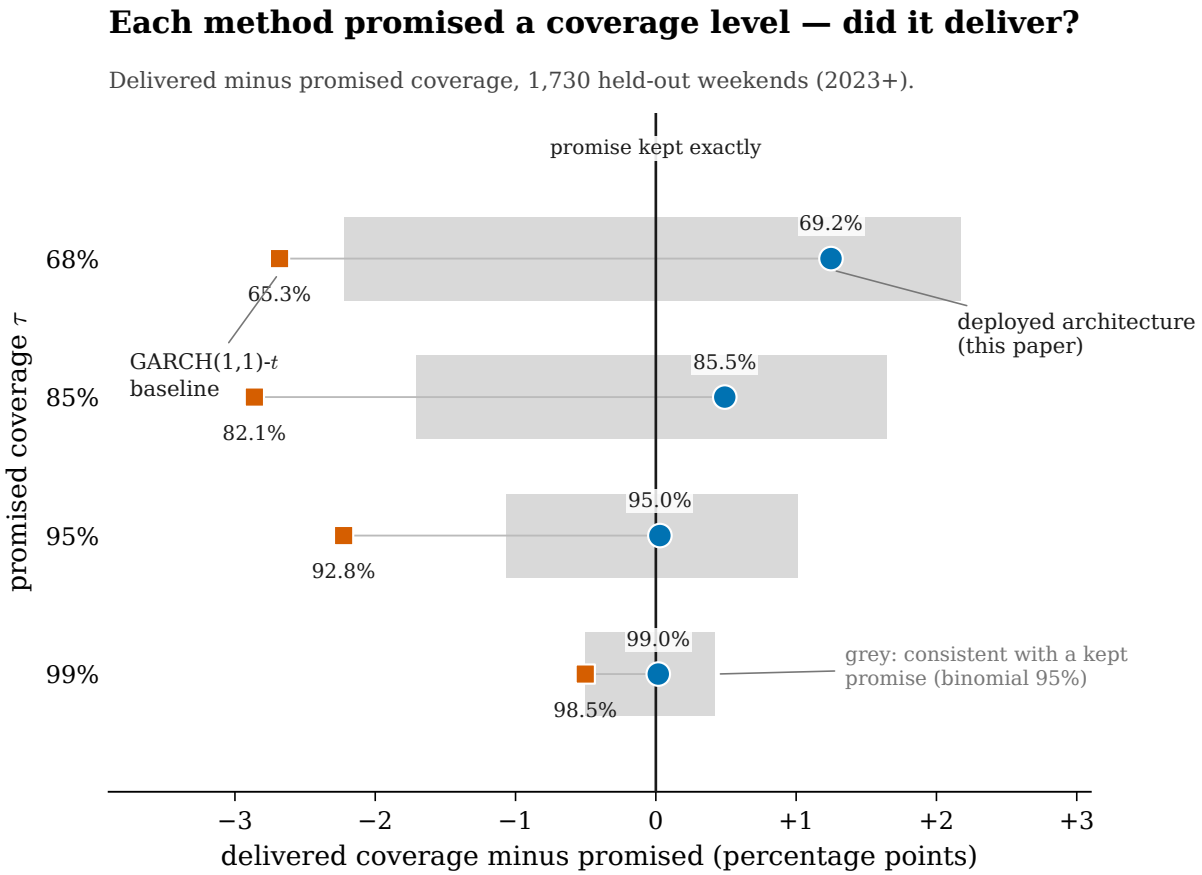

Figure 5: H3. Delivered minus promised coverage at the four served anchors, deployed vs GARCH(1,1)-t, on the 1,730 symbol-weekend (173-weekend) 2023+ OOS slice; the shaded region is the exact binomial 95% acceptance band around each promise. Widened post hoc to matched 95% delivered coverage, GARCH-t's half-width is ≈378 bps against the deployed 370.6 — comparable, but recoverable only after seeing the outcomes, whereas the deployed band stated its coverage in advance. The comparison is restricted to methods that emit a calibrated coverage band — incumbent oracle surfaces (Pyth, Chainlink Data Streams, RedStone, Kamino Scope) do not, so a coverage-vs-sharpness comparison against them is not well-defined (§1.2).

regime variant 19.9%, CI $[-1.5, 35.6]$; $n = 90$), while the short-history refit under-covers (0.91 at $\tau = 0.95$) and its per-regime tail quantile at $\tau = 0.99$ is not identifiable on the window (finite-sample saturation). Per-symbol, the deployment "wins seven of nine" falls to four of nine on matched history — the surviving edge is confined to the thin, volatile perps (GOOGL, HOOD, NVDA, MSTR), while the naive baseline wins the deep-liquid core (SPY, QQQ, GLD), whose perp tracks the underlier tightly enough that a few-months-calibrated conformal band is wider than its empirical quantile. So roughly half the deployment advantage is the calibration record and half is architecture; the architectural half concentrates exactly where a lending consumer's closed-market risk is largest. The architecture is data-hungry by construction — stratifying into regime × per-symbol cells spends observations — which is a genuine limitation (§8) and points to a token-anchored hybrid for the liquid core as future work (§9). Construction detail, the matched-history table, and the small-n power caveat are in Appendix D; none of these numbers enter the master grid — different panel.

6.4 Overnight generalization

If calibration is a property of the method rather than of the weekend, the same architecture should calibrate on the ~17-hour overnight gap. One change — the gap selector admits consecutive-trading-day pairs instead of Friday→Monday pairs — rebuilds the panel as **22,624 overnight rows across 2,412 weeknights** on the same ten tickers, ≈3.8× the weekend panel; every downstream component is reused unchanged. Pooled OOS (6,450 rows × 645 nights) **passes**

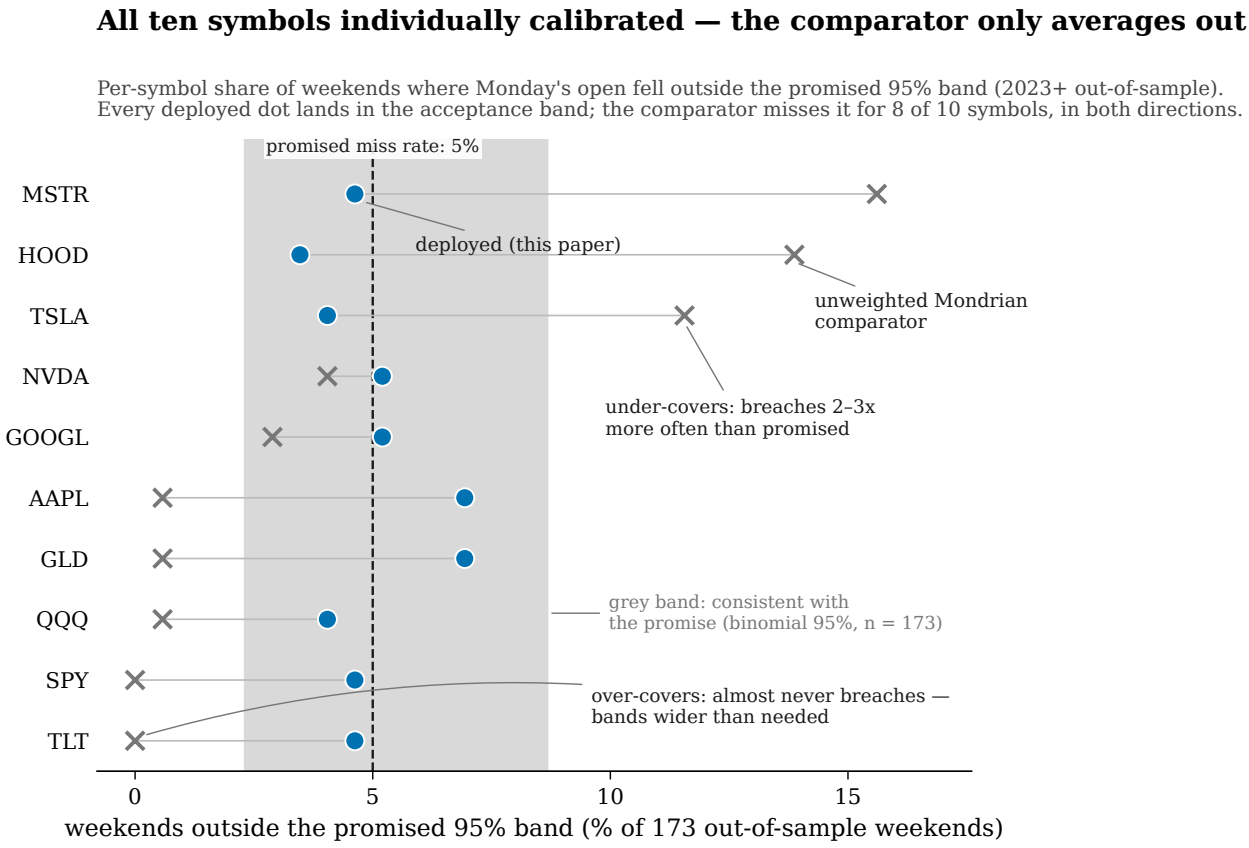

Figure 6: S1. Per-symbol violation rate at $\tau = 0.95$ on the 173-weekend OOS slice — deployed (blue) vs the unweighted-Mondrian comparator (grey ×) — against the binomial 95% acceptance band around the nominal 5%. Deployed: 10 of 10 inside. Comparator: 8 of 10 outside, split between under-coverage on heavy-tail names (MSTR 15.6%, HOOD 13.9%, TSLA 11.6%) and over-coverage on defensive names — pooled parity through compensating per-symbol biases.

Kupiec and Christoffersen at every served anchor, with the OOS-fit $c(\tau)$ collapsing to ≈1.0 — the overnight bands need essentially no post-hoc widening to calibrate. Because consecutive nights are temporally adjacent, we check the i.i.d. strain directly: a moving-block bootstrap on the violation series (block lengths $L \in \{1, 5, 10\}$, 2,000 reps) places the nominal violation rate inside the 95% CI at every τ and every block length, so consecutive-night autocorrelation does not invalidate the coverage claim.

6.5 Earnings nights

`earnings_night` is the one regime the architecture conditions on *a priori*: an earnings release has a publicly known date and session, so the band widens deterministically ahead of the event — calendar-conditioned coverage (§4.7 carries the mechanism). The event study (Fig. H4) shows the served $\tau = 0.95$ half-width widening $\approx 7\times$ exactly at the release gap, against a median realised close→open move at $t = 0$ of $1.8\times$ the baseline half-width — an unwidened band would routinely breach. The fitted earnings-night conformal quantile runs $\approx 8\times$ the normal-night quantile (15.8 vs 2.1 at $\tau = 0.95$; 7.7–$9.3\times$ across anchors); the realised standardised move is correspondingly fat but by a smaller factor ($p_{99} \approx 9.7$ on earnings nights vs ≈ 2.0 on other nights, $\approx 5\times$), and the resulting band is calibrated against that tail, not merely wide: on the OOS slice `earnings_night` realises $0.767/0.967/0.983/1.000$ at $\tau \in \{0.68, 0.85, 0.95, 0.99\}$ on $n = 60$ nights — over-coverage at the upper anchors, the contract-favourable direction on a small cell, though over-wide is still a cost the consumer pays. Two of these cells are honestly *significant* over-coverage, not merely "above the promise": at $\tau = 0.85$ the realised 0.967 against a promised 0.85 is a Kupiec rejection ($p \approx 0.003$, 2 misses where 9 are expected), and the $\tau = 0.95$ cell is borderline — the earnings band is measurably too wide on the current cell, not just conservatively so. The 1.000 cell at $\tau = 0.99$, by contrast, is not distinguishable from exact: at $n = 60$ exact calibration expects only 0.6 misses, so zero is within noise. All three tighten as the earnings panel accumulates; the direction (over-wide) is the safe one for a lending consumer, but it is a disclosed efficiency cost, not a free calibration win.

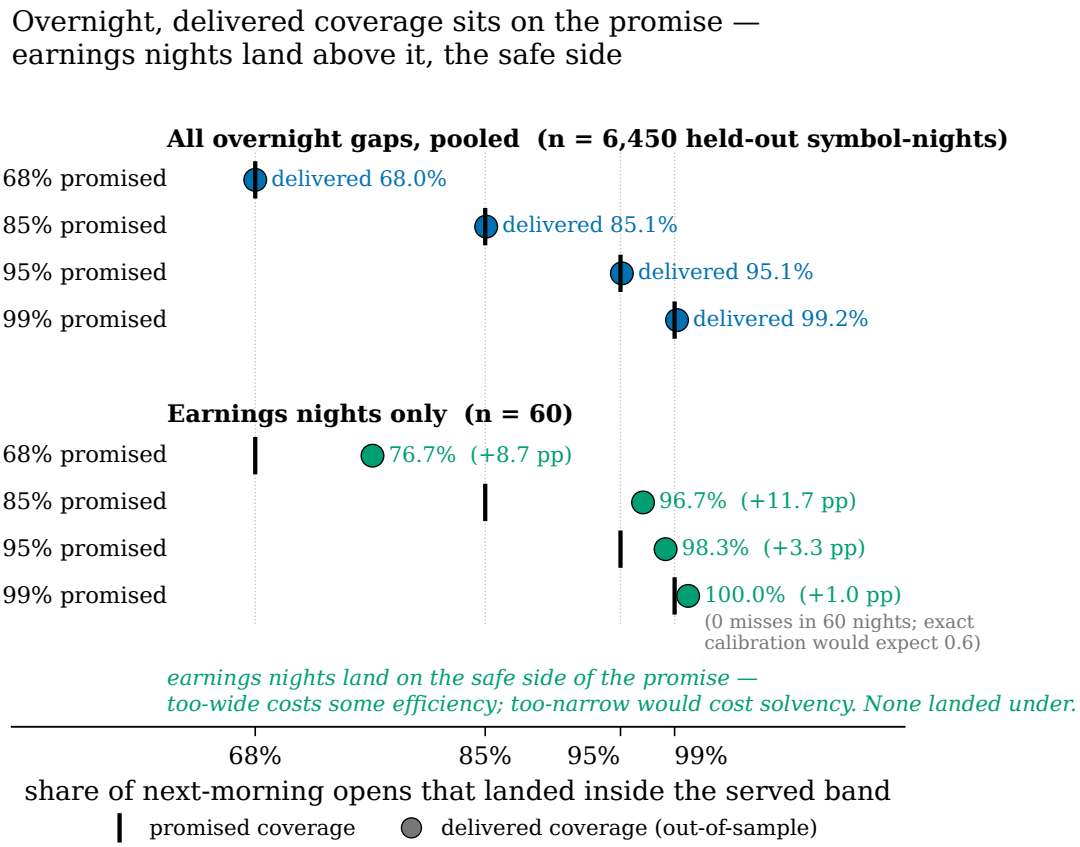

Figure 7: S2. Promised (tick) vs delivered (dot) overnight coverage: pooled ($n = 6,450$ held-out symbol-nights) and earnings nights only ($n = 60$), OOS 2023+. Pooled delivered coverage is 68.0/85.1/95.1/99.2 against 68/85/95/99 promised, with Kupiec unrejected at all four anchors; earnings nights land above the promise at every anchor — the safe (over-wide) side, significantly so at $\tau = 0.85$ (Kupiec $p \approx 0.003$) — and the 99% row's 100.0% is 0 misses in 60 nights where exact calibration would expect 0.6.

The in-sample machinery check, tertile decompositions, full DQ/Berkowitz diagnostics, path coverage, per-asset-class decomposition, and the simulation study are in Appendices B–C; the joint tail belongs to §8. Where the contract fails is §8's subject.

7 What is load-bearing

Three components are load-bearing: per-regime conformal stratification, per-symbol $\hat{\sigma}$ standardization, and the $\hat{\sigma}$ selection procedure. Each is isolated by a comparator that differs from the deployed architecture in exactly that component, and every delta carries a block-bootstrap 95% confidence interval (1,000 resamples, paired by `(symbol, fri_ts)`). Full tables and the supporting robustness battery are in Appendix D.

Stratification, against a constant buffer. The constant-buffer comparator is the alternative a protocol team would reach for first: the Friday close held forward with one global symmetric buffer $b(\tau)$ fit on the training panel. Carried to the 2023+ holdout it under-covers at every $\tau \leq 0.95$ — deficits of -14.2, -12.0, and -5.4 pp at $\tau = 0.68, 0.85, 0.95$, each rejected by Kupiec at $p < 10^{-6}$. The mechanism is non-stationarity: $b(\tau)$ is fit on a 2014–2022 window calmer than the holdout, and a single global scalar has no channel through which to adapt. Even re-fit on the evaluation slice — an oracle-fit, non-deployable configuration that concedes the comparator its most generous trade-off — the deployed architecture is narrower at matched coverage: -5.7% $[-10.6, -0.5]$ at $\tau = 0.85$ and -6.3% $[-12.1, -0.4]$ at $\tau = 0.95$. The stratified surface concentrates its width in `high_vol` ($+52\%$ over the matched buffer), the regime where the constant buffer's violations otherwise cluster.

$\hat{\sigma}$ standardization, against an unweighted Mondrian. This is the architecture's load-bearing calibration result, and — unlike the tokenized head-to-head of §6.2 — it is a matched-history comparison: both arms are fit on the

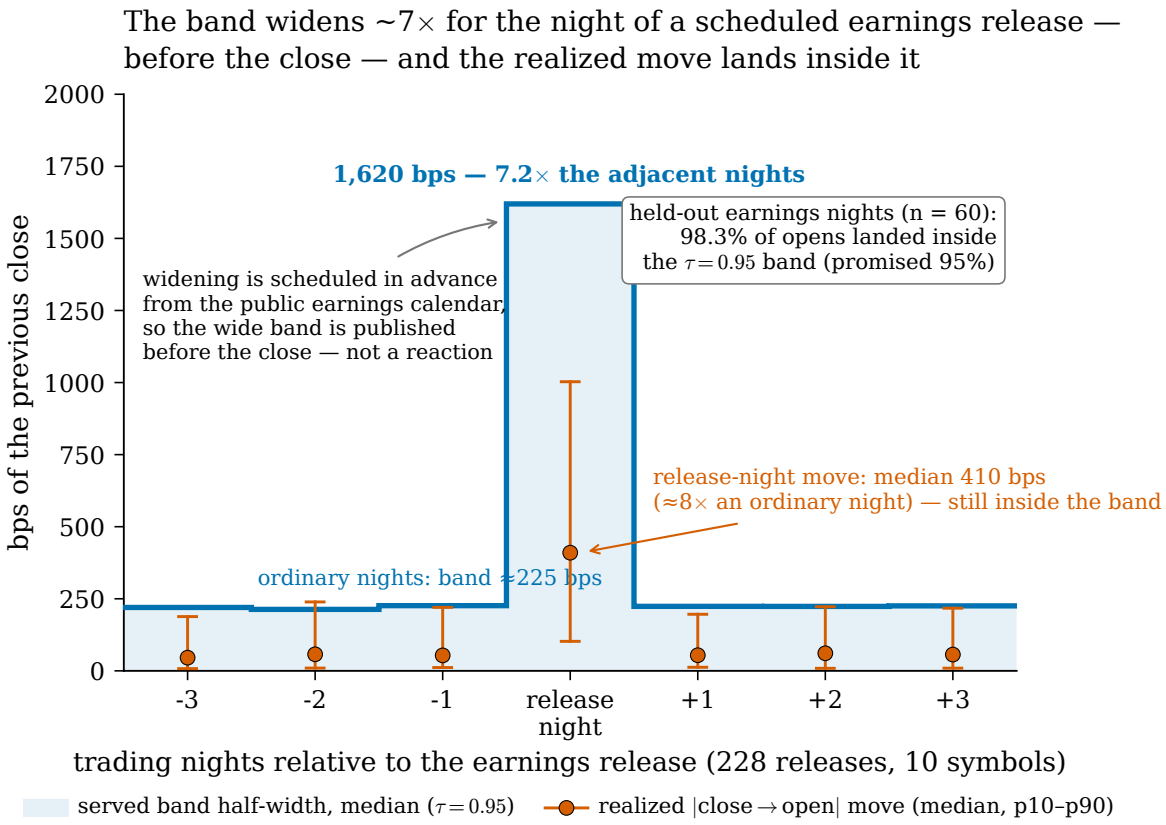

Figure 8: H4. Served $\tau = 0.95$ band half-width (blue step: median across 228 releases — the 229 earnings-night gaps of §5.5 less one release that lacks a full ± 3-night event window, in bps of the previous close) around the earnings-release night, with realised |close→open| moves (dot: median; whiskers: p10–p90): ≈225 bps on ordinary nights, 1,620 bps (7.2×) on the release night — published before the close, from the public earnings calendar. On the $n = 60$ held-out earnings nights, 98.3% of opens landed inside the band (promised 95%). The flat post-event shoulders reflect the $\hat{\sigma}$ de-contamination of §4.7: earnings residuals are excluded from the EWMA scale pool, so subsequent normal nights do not inherit inflated widths.

identical 2014–2026 panel and differ *only* in the $\hat{\sigma}$ standardization, so the effect is architecture, not calibration-sample length. The unweighted-Mondrian comparator is the identical architecture minus $\hat{\sigma}$: same point estimator, regime classifier, split, and finite-sample rank formula, with the conformity score left un-standardized. Both pool to nominal at $\tau = 0.95$ (0.950, Kupiec $p = 0.956$), but the comparator passes per-symbol Kupiec on 2 of 10 symbols; adding $\hat{\sigma}$ standardization takes it to 10 of 10. The comparator's pooled figure is reached through compensating per-symbol biases — heavy-tail tickers under-cover, defensive tickers over-cover, and the cross-sectional average lands on τ by accident. That mechanism is not specific to financial panels, and the obvious remedy is the wrong one: Rafe and Das report the same compensation in complex survey data — nominal marginal coverage carrying roughly 13 pp weighted subgroup gaps — and find that adding group-specific (Mondrian) thresholds made the trade-off worse rather than better [Rafe and Das, 2026]. Partitioning alone does not recover subgroup calibration; standardizing the conformity score *before* the per-regime quantile is fit does, and spends no degrees of freedom on new cells.

Two channels, two axes — against online conformal. Each load-bearing component answers a different conditioning question, and a comparison that measures only one axis will mis-attribute both. We tested this by replacing the calibration engine entirely with the online-conformal family the conformal-VaR literature benchmarks against [Gibbs and Candès, 2021, 2022, Angelopoulos et al., 2023] — identical panel, point estimator, split and metrics, with the online state warmed on the training period so each method sees exactly the data the split-conformal arms are fit on — and by completing the $\{\hat{\sigma}, \text{partition}\}$ two-by-two whose fourth cell ($\hat{\sigma}$ *without* the partition) had never been run. Both experiments were run on the weekend panel and on the overnight panel of §6.8.

On the **per-symbol** axis, $\hat{\sigma}$ standardization is what does the work, and the engine is irrelevant. Pooled on the un-standardized score every engine fails at $\tau = 0.95$ — ACI 3/10, DtACI 3/10, conformal PI 2/10 — reproducing the unweighted-Mondrian failure at 2/10; add $\hat{\sigma}$ and every engine reaches 10/10. Split-conformal on the $\hat{\sigma}$-standardized

score with *no* partition also reaches 10/10. Pooled realized coverage spans 0.947–0.957 across the entire grid, so neither pooled coverage nor per-symbol calibration separates the architecture from an online learner.

On the **per-regime** axis the ordering inverts, and it is decisive. Dropping the partition while keeping $\hat{\sigma}$ passes per-symbol 10/10 and then fails per-regime 1/3 on weekends and 0/3 overnight: it under-covers `high_vol` at 0.926 and over-covers `normal` at 0.965, averaging to nominal by cancellation — the §7.2 compensation mechanism reappearing one level up, along the regime axis instead of the symbol axis. No online configuration recovers it either. Across all twelve online arms, weekend `high_vol` realized coverage spans 0.905–0.932 against the deployed 0.955, and overnight `earnings_night` spans **0.317–0.383 against the deployed 0.983** — a band sized for an ordinary night, breached on roughly two earnings nights in three while claiming $\tau = 0.95$. The deployed architecture is the only configuration tested, on either panel, that passes pooled, per-symbol *and* per-regime simultaneously.

The mechanism is structural rather than a matter of tuning. An online learner adapts a *global* miscoverage level from realized feedback; it converges to the right average and has no channel through which to widen for a *scheduled, identifiable* event before that event is observed. The Mondrian partition is exactly that channel: `earnings_night` is known from the calendar on the afternoon the band is served, and the deployed band widens to 2,508 bps against 342 bps for the same night un-partitioned. This also reprices the sharpness comparison. Per-symbol ACI is 8.7% narrower than deployed on the weekend panel (338.4 vs 370.6 bps) and several online arms are narrower overnight — but that narrowness is bought by under-provisioning precisely the regimes where the risk concentrates, which is not a sharpness gain under a calibration constraint but a failure to meet it conditionally. Full grid in Appendix D. The fix costs a $+4.5\%$ pooled half-width tax ($+15.93$ bps, CI $[+1.05, +30.73]$), which resolves into a redistribution rather than a uniform widening: $+17.8\%$ on equities, -48.8% on the defensive class (GLD, TLT). What that tax buys is stability everywhere the claim is sliced: leave-one-symbol-out realized-coverage dispersion tightens 5.9×, and the per-symbol Berkowitz spread — a test that the full predictive distribution, not just the band edge, is calibrated — contracts from a 250× range to 5.2×.

The selection procedure. Selection was pre-registered: three gates over a five-variant $\hat{\sigma}$ family (a K=26 trailing-window baseline, EWMA half-lives {6, 8, 12}, and a 50/50 blend). Gate 1 requires zero rejections on an 80-cell split-date Christoffersen grid; Gate 2 requires per-symbol Kupiec 10/10 at $\tau = 0.95$; Gate 3 requires the 95% bootstrap CI on width change versus baseline within $+5\%$ at every τ. Gates 1 and 2 do not discriminate: under Benjamini–Hochberg correction — which controls the expected share of false rejections across the 80-cell grid — no variant has any rejected cell, and all five pass Gate 2. Gate 3 is decisive: EWMA HL=8 is -3.83% $[-6.15, -1.88]$ narrower than baseline at $\tau = 0.95$ and significantly narrower at three of four τ anchors, with calibration preserved. A forward-tape harness re-scores all five variants on weekends the selection never saw, under frozen per-variant schedules and with no re-selection: across $N = 11$ forward weekends (2026-05-01 → 2026-07-10) the variants are near-indistinguishable — pooled $\tau = 0.95$ half-widths span 293.6–321.5 bps at identical realized coverage — and the re-validation raises no flag, though N remains below the $N \geq 13$ needed for moderate per-variant power, so this is corroboration rather than confirmation.

Everything else in the ablation battery — the near-identity $c(\tau)$ correction, the quartile-cutoff, split-anchor, and regime-index ablations, the per-symbol ablation figure, and all full tables — is in Appendix D; the head-to-head against a tokenized-tracking baseline, the strongest external comparator, appears in §6.

8 Where it fails

The disclosed boundary. The served bands are per-anchor calibrated (Kupiec passes at every τ; §6) but not full-distribution calibrated. Pooled DQ rejects at $\tau = 0.95$, and pooled Berkowitz rejects on the same fit. The two rejections are one residual, and it localises exactly: the lag-1 signal lives in the cross-sectional within-weekend ordering ($\hat{\rho}_{\text{cross}} = 0.354$, $p < 10^{-100}$, $n = 1,557$), not the temporal within-symbol ordering ($\hat{\rho}_{\text{time}} = -0.032$, $p = 0.18$). Symbols co-breach within a weekend — an equity cluster moving against a safe-haven cluster (GLD, TLT) — not in streaks over time. $\hat{\sigma}$ standardisation operates within a symbol's history, not across symbols within a weekend, so the deployed architecture is structurally unable to reach this residual (Appendix F). It is disclosed, not fixed; the rest of this section measures it.

The joint tail, measured. For each OOS weekend w with full 10-symbol coverage ($n = 173$), let k_w be the number of symbols breaching their τ-band. We compare its empirical distribution to three baselines: $\text{Binom}(10, 1-\tau)$ independence, a Gaussian copula on the empirical correlation matrix, and a Student-t copula on Student-t marginals (Appendix B). At $\tau = 0.95$, $P(k_w \geq 3) = 4.62\%$ [1.73, 7.51] against a binomial 1.15%, and the variance of k_w over-disperses by a uniform $\approx 2.3\times$ vs independence at every served τ (ratios 2.31–2.35) while staying inside the t-copula envelope from below. Mean k_w matches every baseline to 0.003, so the miss is purely in the joint tail: bracketed between independence and a homogeneous t-copula at $\hat{\nu} \approx 6$ (Figure H5).

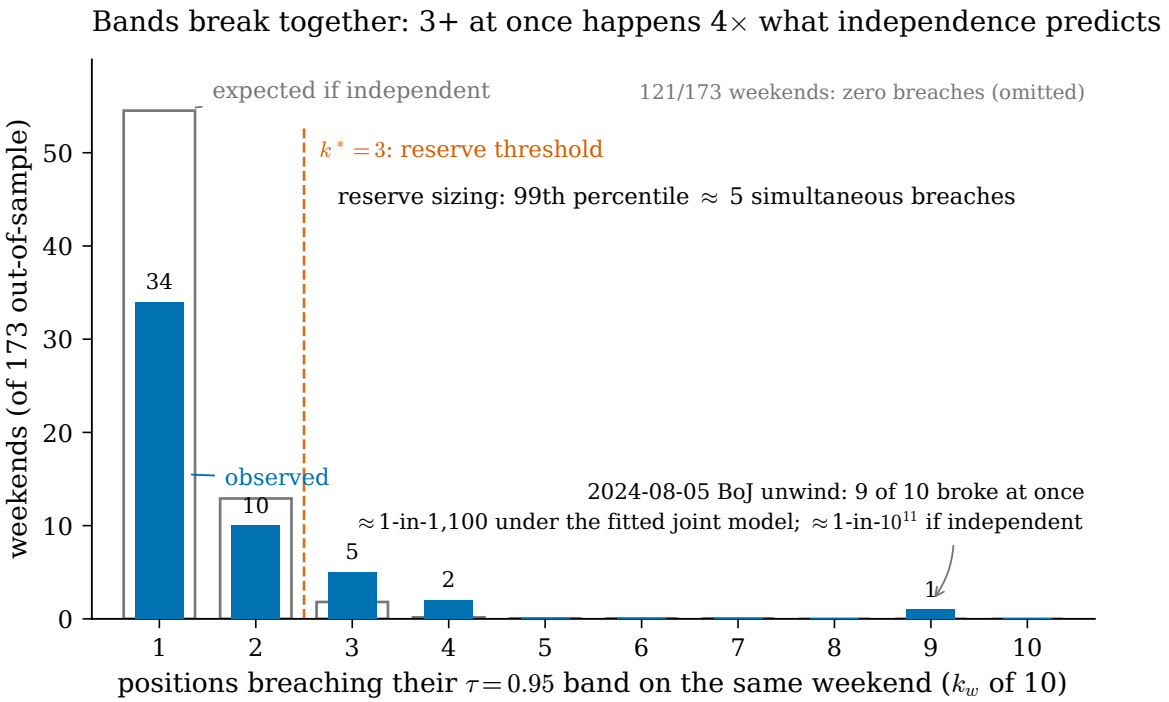

Figure 9: H5. Joint tail of weekend band breaches at $\tau = 0.95$ across the 10-symbol universe, 173 OOS weekends: blue bars count weekends with exactly k_w simultaneous breaches; grey outlines are the expectation under independent breaches, $\mathrm{Binom}(10, 0.05)$; the 121 zero-breach weekends are omitted. Three or more positions breached together on 8 weekends (4.6%, ≈1 in 22) — four times the 1.15% independence predicts. At the $k^* = 3$ threshold, the empirical 99th percentile is ≈5 simultaneous breaches (t-copula ≈6) vs ≈2 under independence. The $k_w = 9$ weekend is the 2024-08-05 BoJ unwind: ≈1-in-1,100 under the fitted t-copula ($\hat{\nu} = 6.04$; appendix figure), ≈1-in-10^{11} under independence.

Reserve guidance. The distribution converts into an operating number: $k^* = 3$ at $\tau = 0.95$ — the smallest joint-breach count whose empirical hit-rate is nearest 5% — with 0/24 Kupiec stability rejections across the 24-cell (τ × convention × sub-period) grid, and an empirical 99th percentile of ≈ 5 simultaneous breaches on a 10-symbol book (binomial: ≈ 2; t-copula: ≈ 6). Reserve to the empirical percentile, not the binomial one. Publishing this distribution is enabled by the receipt design: k_w is computed by re-checking archived receipts against realised opens — an audit any third party can repeat — and a calibration-opaque oracle, with no stated coverage target to breach, has no defensible basis to publish a joint-breach distribution at all. No incumbent oracle (Pyth, Chainlink Data Streams, RedStone, Kamino Scope) publishes this distribution.

The worst observed weekend. Friday 2024-08-02 → Monday 2024-08-05, the BoJ rate-hike yen-carry unwind: $k_w = 10$ at $\tau = 0.85$, 9 at $\tau = 0.95$ (only TLT inside), 5 at $\tau = 0.99$. The per-symbol scaling held — MSTR's $\tau = 0.85$ half-width was 663 bps against SPY's 85 — but nine symbols opened lower in concert by 4–27 $\hat{\sigma}$-scale returns, and the miss was common-mode: no per-symbol band, however well calibrated, absorbs a joint move of that shape. The operational read: a consumer monitoring k_w saw 10 at $\tau = 0.85$ by the Monday print, double the fitted $k^* = 5$ at that anchor. At $\tau = 0.85$ (where $k_w = 10$) the fitted t-copula puts this weekend at ≈ 1-in-475; independence puts the same all-ten-breach event at $0.15^{10} \approx$ 1-in-1.7×10^8 — the gap between those figures is the residual this section names. No method on the comparison grid attains its nominal coverage at $\tau \leq 0.95$ on this weekend (Appendix B, M6 re-run).

Remaining bounds. Each bounds a claim; none reverses one.

- *Shock conditioning.* Conditional on a shock realised move ($|z| \geq 0.67$ on the realised-move z-score, the shock tertile of B.5; $n = 632$), coverage at nominal $\tau = 0.95$ is 87.8%; half-width is nearly flat across tertiles (365–380 bps), so the band does not widen under shock — it misses more often, via the same common-mode channel.
- *Stationarity.* The guarantee assumes the standardised score's conditional distribution is stationary. Three violations no backtest grid can pre-empt: structural breaks beyond the 2014–2026 panel, regime-labeler drift (the trailing VIX percentile shifts endogenously), and factor-switchboard drift (a second MSTR-style pivot).

- *Asymmetry.* The documented failure modes are one-sided toward over-coverage — a sharpness cost, not a contract breach. Under-coverage is the alarm side, where the k_w monitor points.
- *Endpoint, not path.* On a 24/7 perp slice ($n = 118$), the $\tau = 0.95$ endpoint band under-covers the intra-weekend path by +16.1 pp; the contract remains an endpoint claim, and the gap approximately closes at $\tau = 0.99$.
- *Underlier, not xStock.* The band predicts the underlier's open, not the token's price. A tokenized-side conditional-mean baseline exists (Cong et al.'s $\lambda = 0.903$ passthrough); as deployed the band beats it substantially on Winkler, but on matched calibration history that edge narrows to $\approx 20\%$ with a CI including zero (§6.2, Appendix D) — most of the deployment margin is the 12-year record, not the architecture.
- *Data-hungry by construction.* The architecture stratifies into regime × per-symbol cells and fits a per-symbol scale, which spends observations: on a short (few-month) calibration window the per-regime conformal quantile saturates (the $\tau = 0.99$ tail is not identifiable) and the refit under-covers (0.91 at $\tau = 0.95$; Appendix D). The deployed calibration relies on the full 2014–2026 panel, and thinly-traded new symbols inherit weaker per-cell calibration until their history accumulates.
- *Off-anchor τ.* Only the four anchors $\{0.68, 0.85, 0.95, 0.99\}$ are audited; intermediate τ is served by interpolation, and the receipt exposes the interpolated multipliers, so the case is detectable from the receipt alone.
- *Earnings nights.* The `earnings_night` cell carries small OOS n (≈ 60) and currently over-covers.
- *Location shifts.* In simulation, $\hat{\sigma}$ absorbs a pure mean jump as variance; a deployed CUSUM monitor closes the detection gap (Appendix C).

k^* is reserve guidance under the deployed schedule, not an optimal policy — that would need a protocol cost model, out of scope (§3.5). Every number here is re-derivable from archived receipts and realised opens: the paper can say where its bands fail with the same precision it says where they hold.

9 Conclusion

This paper's contribution is an interface and an architecture that delivers it. The interface: target coverage τ is the consumer-chosen input, the band is the output, and every read carries a receipt any third party can re-derive from public data. The architecture: locally-weighted Mondrian split-conformal, whose held-out calibration — leave-one-symbol-out coverage 0.9497 ± 0.0128 at $\tau = 0.95$, 40/40 Kupiec on the symbol × τ grid against GARCH-t's 31/40 — rests on three load-bearing components isolated in §7 and corroborated on synthetic ground truth in Appendix C.

The primitive stands independent of this architecture. Nothing in the receipt contract requires split-conformal: any engine — fully conformal, or not conformal at all — that can populate the calibration surface of §3 can serve the same τ-in, band-plus-receipt-out contract and inherit its auditability.

The advantage has an honest scope: against the tokenized-tracking baseline the architecture ties on SPY and TSLA — the deepest perps and largest collateral — and wins on the thin-liquidity long tail, where a consumer's closed-market risk need is greatest. The value concentrates where incumbents are weakest.

The contract, as served, is complete for its stated scope. A consumer today chooses τ against the four audited anchors, verifies any read against public data, and wires the receipt directly into collateral policy — the band into loan-to-value, the published joint tail into reserves ($k^* = 3$ at $\tau = 0.95$ for a ten-name portfolio) — and can hold the provider to the realised rate on every future window. A full-distribution variant that removes §8's boundary is already characterised in Appendix F, where a cluster-conditional correction moves the pooled DQ p-value from rejection to 0.912 and *tightens* half-width $\approx 11.6\%$ at $\tau = 0.95$, with the benefit bounded to the $\tau \geq 0.95$ anchors where it is calibrated (at lower τ endpoint scoring on the unpartialled residual remains dominant; §F.10); it is not deployed, and nothing this paper claims depends on it. A one-sided instantiation for the lending track is characterised in Appendix G: because a collateral holder is exposed in one direction only, the deployed two-sided band at τ hands them $(1 + \tau)/2$ protection, and serving the downside quantile directly frees roughly a third of the collateral buffer at the $\tau = 0.85$ deployment default while passing Kupiec at every anchor on both panels — where imposing symmetry fails it at three of four overnight anchors. It is not deployed and carries none of §6's held-out battery. Two further architectural extensions the evidence points to directly: a token-anchored hybrid that leans on the 24/7 tokenized price for deep-liquid names — where §6.2's matched-history test shows a naive tracker already competitive — while reserving the conformal-on-underlier machinery for the thin, volatile tail where it wins; and a shorter-history variant that pools regime cells to stay identifiable before a symbol's per-cell calibration accumulates. The §7 comparison against the online-conformal family locates the contribution precisely. Those methods match the architecture on pooled and per-symbol calibration once given a per-symbol channel, and several are narrower — but none of them delivers *conditional* coverage: across every online arm, overnight `earnings_night` coverage lands between 0.317 and 0.383 against a claimed 0.95, where the deployed

partition realizes 0.983. An adaptive learner converges to the right average and has no channel through which to widen for a scheduled event it has not yet observed. Sharpness bought by under-provisioning the regimes that carry the risk is not sharpness under a calibration constraint. One asymmetry between the two closed windows shapes what this work is for. Extended-hours reform targets the overnight gap and not the weekend (§1.4): every filed proposal is five-day, so a 24/5 market removes the overnight window and leaves the weekend one intact. The weekend result is therefore the durable contribution, and the overnight result is both the generality test (§6.8) and the surface with the shorter horizon. The two are also beginning to diverge under measurement — they carry different regime taxonomies already, and characterised extensions suggest they call for different per-cell treatment and, for the overnight profile, a serving contract that trades the frozen artefact's pre-commitment for a published, checkpointed adaptive state. Whether the primitive should be instantiated identically across closed windows, or tuned per window with the receipt disclosing which contract is in force, is the open design question this evidence raises. Path-aware truth, decision-theoretic τ selection, multi-region replication, and other RWA classes are out of scope.

The empirical bottom line is one number and one boundary. The number: 0.9497 ± 0.0128 held out by symbol, 40/40 held out across the grid. The boundary: pooled DQ rejects — the bands are per-anchor calibrated, not full-distribution calibrated (§8). We have treated that boundary the way an oracle should treat its own uncertainty: measured (§8's joint-tail distribution), priced (the k^* reserve threshold), monitored (the forward-tape CUSUM bank, Appendix B), and disclosed on every read. The residual a consumer inherits is no longer implicit in a methodology document — the calibration statement becomes a measurable object on the wire.

A Reproducibility

This appendix consolidates everything needed to reproduce the §6 empirical results, the §7 ablation, and the Appendix B / Appendix D robustness diagnostics from the publicly-available source data plus the Soothsayer repo. The deployment is a five-line lookup against sixteen pre-fit scalars; reproducibility is determined by the panel construction, the schedule fits, and the serving formula — all of which are deterministic given the panel.

A.1 Algorithm boxes

We give pseudocode for three procedures: the M6 fit (deployment-artefact build), the M6 serve (runtime band lookup), and the strictly-pre-Friday per-symbol $\hat{\sigma}$ EWMA construction. §4 cites Algorithms 1–2 as the recipe; nothing there is duplicated here beyond the pseudocode itself. Implementations: Python at `src/soothsayer/{backtest/calibration.py,oracle.py}`, Rust port at `crates/soothsayer-oracle/src/{config,oracle,types}.rs` (currently mirrors the M5 reference path; M6 wire-format slot reserved as `forecaster_code = 3`).

Algorithm 1 — M6 fit (one-shot, produces the 16 deployment scalars; `scripts/build_lwc_artefact.py`).

```
Input:  panel D = {(s, t, p_Fri, mon_open, r_F, r) : i = 1, ..., N}
        split date d_split = 2023-01-01
        anchor τ-grid T = {0.68, 0.85, 0.95, 0.99}
        c-grid C = {1.000, 1.001, ..., 5.000}
        EWMA half-life HL = 8 weekends
        warm-up minimum past_obs_min = 8

Output: sigma_hat_table  sigma_hat_s(t)         per-(symbol, fri_ts) parquet column
        quantile_table   q_r(τ)         dim 3 × 4 = 12 scalars
        c_bump_schedule  c(τ)           dim 4 scalars
        delta_shift_schedule  δ(τ)      dim 4 scalars (identically zero under M6)

1.  for each row i ∈ D:
        p_hat_i ← p_Fri,i · (1 + r_F,i)             # factor-adjusted point
        rrel_i ← (mon_open_i − p_hat_i) / p_Fri,i   # signed relative residual
2.  for each (symbol s, weekend t) ∈ D:
        prior ← {rrel_{i'} : symbol(i') = s, t' < t}    # strictly pre-Friday
        if |prior| < past_obs_min:
            mark row as warmup_drop and skip
        else:
            sigma_hat_s(t) ← EWMA(prior, half_life = HL)        # weekend half-life decay
3.  D_train ← {i ∈ D : t_i < d_split, not warmup_drop}
    D_oos   ← {i ∈ D : t_i ≥ d_split, not warmup_drop}
4.  for each row i in D_train ∪ D_oos:
```

```
        score_i ← |mon_open_i − p_hat_i| / (p_Fri,i · sigma_hat_{s_i}(t_i))  # std.
5.  for each (regime r, target τ) ∈ {normal, long_weekend, high_vol} × T:
        S_r ← {score_i : i ∈ D_train, r_i = r}
        k ← ⌈τ · (|S_r| + 1)⌉                    # finite-sample CP rank
        q_r(τ) ← sort(S_r)[k]               # trained standardised quantile
6.  for each τ ∈ T:
        c(τ) ← smallest c∈C with mean_{D_oos} 1{score_i ≤ q_{r_i}(τ)·c} ≥ τ
7.  δ(τ) ← 0 at every anchor under M6 (walk-forward sweep; per-symbol sigma_hat
           tightens cross-split coverage variance, so no shift is needed)
8.  return (sigma_hat_s(t), q_r(τ), c(τ), δ(τ))
```

Algorithm 2 — M6 serve (runtime band; `Oracle.fair_value_lwc` in `src/soothsayer/oracle.py`).

```
Input:  symbol s, as-of date t, target coverage τ ∈ [0.68, 0.99]
        lookup row (p_Fri, regime r, sigma_hat_s(t)) from per-Friday parquet at (s, t)
        deployment scalars (q_r(τ), c(τ), δ(τ) ≡ 0) from Algorithm 1

Output: PricePoint{ τ, δ(τ), p_hat, lower, upper, r, c, q_eff, q_r, sigma_hat }

1.  c_eff ← interpolate c on the τ-grid at τ
2.  q_eff ← c_eff · q_r(τ)              # standardised quantile (sigma_hat below)
3.  p_hat ← p_Fri · (1 + r_F)              # factor-adjusted point (centre)
4.  half ← q_eff · sigma_hat_s(t) · p_Fri      # price-unit half-width
5.  lower ← p_hat − half;   upper ← p_hat + half
6.  return PricePoint(τ, δ(τ) = 0, p_hat, lower, upper, r,
                      diagnostics{c_eff, q_eff, q_r(τ), sigma_hat_s(t)})
```

Algorithm 3 — Per-symbol $\hat{\sigma}$ EWMA construction (`scripts/build_lwc_artefact.py` step 2; soft constants `SIGMA_HAT_HL_WEEKENDS = 8`, `SIGMA_HAT_MIN = 8`).

```
Input:  panel D, EWMA half-life HL = 8 weekends, past_obs_min = 8

Output: per-(symbol, fri_ts) sigma_hat_s(t) column on the artefact parquet

1.  λ ← 0.5 ** (1.0 / HL)        # ≈ 0.917 per past Friday
2.  for each symbol s in D:
        rows_s ← rows of D with symbol = s, sorted by fri_ts
        for each row i in rows_s:
            prior ← {rrel_j : j∈rows_s, fri_ts(j) < fri_ts(i)}  # pre-Fri
            if |prior| < past_obs_min:
                sigma_hat_s(fri_ts(i)) ← undefined  (row marked warmup_drop)
            else:
                weights ← [λ ** (|prior| − k − 1) for k in 0..|prior|]
                sigma_hat_s(fri_ts(i)) ← weighted_std(prior, weights)
3.  return sigma_hat column
```

The $\hat{\sigma}$ rule itself was selected from a five-variant ladder under the pre-registered three-gate criterion of §7 (full procedure and tables in Appendix D); alternative variants remain buildable for archival reproduction via `scripts/build_lwc_artefact.py --variant {baseline_k26, ewma_hl6, ewma_hl12, blend_k26_ewma_hl8}`.

A.2 Hyperparameter specification

The M6 deployment artefact is fully specified by the 16 scalars below plus four "soft" pipeline constants and the $\hat{\sigma}$ rule constants. All values are recorded in `data/processed/lwc_artefact_v1.json` (the audit-trail sidecar) and loaded at module import into `src/soothsayer/oracle.py:LWC_REGIME_QUANTILE_TABLE` (with the schedule constants hard-coded alongside); the Rust unit test `lwc_constants_match_sidecar` enforces ≤ 2 ULP agreement between the hard-coded constants and the sidecar on every refit (A.8).

Trained per-regime quantiles $q_r(\tau)$ on standardised residuals (12 scalars, fit on the 4,186-row pre-2023 train set after the 80-row $\hat{\sigma}$ warm-up drop):

regime (n)	$\tau = 0.68$	$\tau = 0.85$	$\tau = 0.95$	$\tau = 0.99$
normal	0.7767	1.2195	1.9682	3.3280
long_weekend	0.8638	1.3649	2.2208	4.0152
high_vol	1.1086	1.9433	3.0971	6.4563

Note these are *standardised* quantiles (dimensionless; multiplied at serve time by $\hat{\sigma}_s(t)$ to yield a relative-residual quantile in basis points). The high_vol $\tau = 0.99$ value of 6.46 — a 6.46-σ shock at the panel-relative scale — is the headline-tail shape M6 is calibrated to absorb.

OOS-fit $c(\tau)$ multiplicative bumps (4 scalars, fit on the 1,730-row 2023+ OOS via Algorithm 1 step 6):

τ	0.68	0.85	0.95	0.99
$c(\tau)$	1.000	1.000	1.079	1.003

Three of four near-identity; only $c(0.95) = 1.079$ carries meaningful OOS information. §8 carries the provenance disclosure for that one scalar; B.7 carries the detail.

Walk-forward $\delta(\tau)$ shifts (4 scalars; identically zero under M6 — per-symbol $\hat{\sigma}$ standardisation tightens cross-split realised-coverage variance enough that no shift is required):

τ	0.68	0.85	0.95	0.99
$\delta(\tau)$	0.000	0.000	0.000	0.000

The δ schedule is retained as a 4-zero vector in the artefact JSON for shape-compatibility with the receipt schema.

Soft constants (pipeline shape; not OOS-tuned):

Constant	Value	Source
Train/OOS split anchor	2023-01-01	§5.6; split-date sensitivity in B.6
Conformity score	$\lvert y - \hat{p}\rvert / (p_{\text{Fri}} \cdot \hat{\sigma}_s)$	§4.4 (standardised)
Point estimator	$p_{\text{Fri}} \cdot (1 + r_t^F)$	§4.2 / §5.4 (factor switchboard)
Factor switchboard	per-symbol $\rho(s,t)$	§5.4, MSTR pivot 2020-08-01
Regime classifier	3-bin $\rho \in \{\text{normal}, \text{lw}, \text{hv}\}$	§5.5
`c-grid` resolution	$\Delta c = 0.001$	Algorithm 1 step 6
Off-grid τ interpolation	linear on the four anchors	§4.6
$\hat{\sigma}$ EWMA half-life	8 weekends	§4.3; `SIGMA_HAT_HL_WEEKENDS`
$\hat{\sigma}$ warm-up minimum past obs	8	§4.3; `SIGMA_HAT_MIN`

A.3 Data and code availability

Data. All upstream data is publicly available (Yahoo Finance daily bars, CME 1m futures, VIX/GVZ/MOVE, BTC-USD, Yahoo earnings calendar, Kraken Futures perp tape, Pyth Hermes archive, Chainlink Data Streams archive). Soothsayer consumes pre-fetched parquet from the sibling `scryer` repo at `SCRYER_DATASET_ROOT`. The decade-scale weekend panel `data/processed/v1b_panel.parquet` (5,996 rows × 35 columns, 2014-01-17 → 2026-04-24) is built by `scripts/run_calibration.py` calling `soothsayer.backtest.panel.build()`; the overnight panel `data/processed/overnight_panel.parquet` (22,624 rows, 2014-01-16 → 2026-04-23) is built by `scripts/build_overnight_panel.py` calling the same `build()` with `gap_mode="overnight"` plus `yahoo/earnings/v2` session timing and the `yahoo/corp_actions/v1` ex-dividend adjustment (§5.2). All `_fetched_at` cutoffs are preserved in the parquet metadata.

Panel row schema. Each row is a single (s,t) prediction window carrying: `fri_close`, `mon_open`, `gap_days`, `fri_vol_20d` (rolling 20-trading-day std of daily log-returns at the closing print), `factor_ret` (closed-window return of the per-symbol conditioning factor), `vol_idx_fri_close`, `earnings_next_week` (Yahoo `earnings_dates` flag;

session-timed per gap on the overnight panel), and `is_long_weekend` (`gap_days` ≥ 4). §5.2 carries the panel counts; §5.5 the regime shares.

Repository. Code at `https://github.com/ACNoonan/soothsayer`. Results were produced from the `main` branch as of 2026-05-05; the camera-ready snapshot is tagged to the commit covering the corrected overnight battery (2026-06-25) and the $N = 11$ forward tape (2026-07-14). License: see repository LICENSE file.

Layout.

Layer	Path	Purpose
Panel build	`src/soothsayer/backtest/panel.py`	weekend-pair assembly, regime tag, factor join
Calibration helpers	`src/soothsayer/backtest/calibration.py`	`compute_score`, `train_quantile_table`, `fit_c_bump_schedule`, `serve_bands`, `fit_split_conformal_forecaster`
Metrics	`src/soothsayer/backtest/metrics.py`	Kupiec, Christoffersen, Berkowitz, DQ, CRPS, PIT
Python serve	`src/soothsayer/oracle.py`	`Oracle.fair_value_lwc` (M6 path); `Oracle.fair_value` (M5 reference)
Rust serve (parity)	`crates/soothsayer-oracle/`	byte-for-byte both forecasters (`Forecaster::Mondrian`, `Forecaster::Lwc`); 180/180 parity vs Python
On-chain publisher	`programs/soothsayer-oracle-program/`	Anchor program, `PriceUpdate` Borsh wire
M6 artefact build	`scripts/build_lwc_artefact.py`	Algorithm 1
M6 freeze	`scripts/freeze_lwc_artefact.py`	content-addressed freeze for forward-tape eval
M5 artefact build (reference)	`scripts/build_mondrian_artefact.py`	Appendix D ablation rung
Cross-verification	`scripts/verify_rust_oracle.py`	180-case Python ↔ Rust parity probe (90 M5 + 90 M6 LWC)
Robustness battery	`scripts/run_v1b_*.py`, `scripts/run_*.py`	per-symbol diagnostics, vol-tertile, GARCH-N + GARCH-t, split sensitivity, LOSO, per-class, path-fitted
Paper-strengthening runners	`scripts/run_portfolio_clustering.py`, `scripts/run_subperiod_robustness.py`, `scripts/run_v1b_garch_baseline.py --dist t`, `scripts/run_per_symbol_kupiec_all_methods.py`, `scripts/run_kw_threshold_stability.py`	joint-tail clustering (§8 / B.8), sub-period stability (B.6), GARCH-t baseline (B.12), 4-method per-symbol grid (§6.2), k_w threshold stability (B.8)

Forward tape | `scripts/collect_forward_tape.py`, `scripts/run_forward_tape_evaluation.py` | weekly held-out evaluation against the frozen artefact (A.9) |
Figures | `scripts/build_paper1_figures.py` (appendix figures), `scripts/build_fig_h{1,2,3,4,5}.py`, `scripts/build_fig_s{1,2}.py` (main-text figures) | per-figure provenance in A.6 |

A.4 Compute environment

Component	Version	Notes
Python	3.12.13	enforced via `pyproject.toml` `requires-python = ">=3.12,<3.13"`
Environment manager	`uv`	lockfile `uv.lock` checked in
Key deps (Python)	numpy $\geq$ 1.26, pandas $\geq$ 2.2, scipy $\geq$ 1.13, polars $\geq$ 0.20, pyarrow $\geq$ 16.0, statsmodels $\geq$ 0.14, arch (for the B.12 GARCH baselines)	full list in `pyproject.toml`

Component	Version	Notes
Rust	stable (1.81+)	matches Solana toolchain
Anchor	0.30.x	for the on-chain publisher
OS	macOS Darwin 25.4 (development); deployment is platform-agnostic	

A.5 Reproducing the paper's numerical content

This is the single consolidated command listing; §4.8, §5.7, and Appendix E refer here rather than repeating commands. The end-to-end pipeline from raw panel to all paper artefacts:

```
# 0. Environment.
uv sync

# 1. Build the weekend panel from scryer parquet (one-time, ~3 min;
# end-to-end backtest: under 15 min cold, under 1 min warm).
uv run python scripts/run_calibration.py

# 2. Fit M6, write the artefact (Algorithm 1) + sidecar; freeze; smoke-test.
uv run python scripts/build_lwc_artefact.py    # artefact + JSON sidecar
uv run python scripts/freeze_lwc_artefact.py   # content-addressed freeze
uv run python scripts/smoke_oracle.py --forecaster lwc   # cross-regime API demo

# 3. Verify Python <-> Rust parity (A.8); build the Rust serving CLI.
PYTHONUNBUFFERED=1 uv run python scripts/verify_rust_oracle.py
uv run python scripts/build_mondrian_artefact.py    # M5 reference (App D rung)
cargo build --release -p soothsayer-publisher
soothsayer fair-value --symbol SPY --as-of 2026-04-24 --target 0.85
./scripts/deploy_devnet.sh                     # devnet publish (App E)

# 4. Path-coverage diagnostic (B.14) + excursion-inflation λ.
uv run python scripts/run_path_coverage.py
uv run python scripts/proto_excursion_inflation.py

# 4b. Off-hours: overnight panel + calibration battery (§6.4, B.15).
uv run python scripts/build_overnight_panel.py   # overnight panel (§5.2)
uv run python scripts/build_overnight_artefact.py   # overnight artefact + battery

# 5. §6 / §7 / Appendix B / Appendix D robustness battery.
uv run python scripts/run_v1b_per_symbol_diagnostics.py
uv run python scripts/run_v1b_vol_tertile.py
uv run python scripts/run_v1b_garch_baseline.py          # GARCH-Gaussian
uv run python scripts/run_v1b_garch_baseline.py --dist t # GARCH-t (B.12)
uv run python scripts/run_v1b_split_sensitivity.py
uv run python scripts/run_v1b_loso.py
uv run python scripts/run_v1b_per_class.py
uv run python scripts/run_v1b_path_fitted_conformal.py
uv run python scripts/run_v1b_tokenised_tracking_baseline.py  # §6.2, App D

# 6. Joint-tail / stability runners (§8, B.6, B.8).
uv run python scripts/run_portfolio_clustering.py     # joint-tail dist.
uv run python scripts/run_subperiod_robustness.py     # sub-period stability
uv run python scripts/run_per_symbol_kupiec_all_methods.py  # 4-method grid
uv run python scripts/run_kw_threshold_stability.py   # reserve threshold

# 7. Build the paper figures (A.6).
uv run python scripts/build_paper1_figures.py
uv run python scripts/build_fig_h1.py
uv run python scripts/build_fig_h2.py
uv run python scripts/build_fig_h3.py
uv run python scripts/build_fig_h4.py
uv run python scripts/build_fig_h5.py
```

```
uv run python scripts/build_fig_s1.py
uv run python scripts/build_fig_s2.py

# 8. Compile this paper.
cd research/coverage-inversion/build && uv run python build.py --pdf
```

`run_calibration.py` materialises `data/processed/v1b_bounds.parquet`, the per-symbol and pooled calibration surfaces, and refreshes the OOS-evaluation tables; the two overnight scripts materialise the overnight panel and its calibration battery. Forward-tape harness commands are in A.9.

A.6 Per-figure data provenance

Rebuilt for the H1–H5 / S1–S2 figure plan. H1–H3 are new builds; H4, H5, S1, and S2 are redesigns of earlier exhibits (fig11, fig10, fig4, and fig8 respectively); the remaining appendix figures carry their original provenance.

Status (new build / redesign of a prior exhibit / existing) is given in the paragraph above; the table below records each figure's builder and data source. Paths under `data/processed/` and `reports/tables/` are shown without their directory prefix; the `figN_*()` builders are functions in `scripts/build_paper1_figures.py`.

Figure (placement)	Script	Source data (parquet / CSV)
H1a week timeline (§1); H1b gap distributions (§1.1)	`build_fig_h1.py`	`v1b_panel`, `overnight_panel`
H2 anatomy of a read (§3)	`build_fig_h2.py`	`lwc_artefact_v1` (exemplar; else diagrammatic)
H3 promised-vs-delivered coverage (§6.2)	`build_fig_h3.py`	`m6_pooled_oos`, `m6_lwc_robustness_garch_t_baseline`
H4 earnings event-time band (§6.5)	`build_fig_h4.py`	`overnight_panel`, `overnight_artefact_v1`
H5 joint-tail k_w counts (§8)	`build_fig_h5.py`	`paper1_a3_joint_baseline_kw_distribution`
Fig. 1 serving pipeline (§4)	`fig1_pipeline()`	diagrammatic
S1 per-symbol promise audit (§6.2)	`build_fig_s1.py`	`m6_lwc_robustness_per_symbol`, `m6_per_symbol_kupiec_4methods`
S2 overnight promise audit (§6.4)	`build_fig_s2.py`	overnight panel + artefact (§5.2, B.15)
fig6 path coverage (B.14)	`fig6_path_coverage()`	`path_coverage_perp_per_row` + artefact bands
fig7b per-symbol ablation (App D)	`fig7b_oos_ablation()`	`paper1_fig7b_per_symbol_ablation` (per split anchor)
fig9 BoJ band anatomy (B.9)	`fig9_boj_anatomy()`	`lwc_artefact_v1`, `v1b_panel`
simulation figure (App C)	`run_simulation_study.py`	`sim_per_symbol_kupiec`

Former fig0, fig2, and fig5 are superseded (fig0 upgraded into H1; fig2 + fig5 merged into H3) and are no longer emitted into the build.

A.7 Determinism and randomness

The M6 fit and serve paths are deterministic given the panel: no random initialisation, no Monte Carlo, no bootstrap inside the fit. The 4-split powered walk-forward (B.6, Appendix D) uses fixed split fractions {0.4, 0.5, 0.6, 0.7} (the 0.2 / 0.3 fractions are excluded as under-powered for the 4-scalar $c(\tau)$ fit; Appendix D); the four split-date sensitivity anchors (B.6) are fixed at {2021-01-01, 2022-01-01, 2023-01-01, 2024-01-01}; LOSO (§6.1) iterates the ten symbols in lexicographic order. The block-bootstrap CIs reported in §8, B.8, and Appendix D use NumPy's default RNG with seed 0 over 1,000 weekend-block resamples. The simulation study (Appendix C) uses NumPy's default RNG with seed 0 over 100 replications per DGP.

The GARCH(1,1) baselines (B.12) are fit per-symbol via `arch_model(..., dist={"normal", "t"}).fit(disp="off")`; the optimisation is deterministic up to BFGS tolerance (default `tol=1e-8`). Under GARCH-t, NVDA's t-MLE hits the optimizer lower bound $\hat{\nu} = 2.50$ (adjacent to the $\nu \leq 2$ variance-undefined region) and falls back to Gaussian; the fallback is recorded in the `dist_used` column of the output CSV and is deterministic across reruns.

A.8 Independent verification (Python ↔ Rust ↔ on-chain)

The deployed serving stack has three independent implementations of both forecasters that must agree byte-for-byte:

1. **Python reference.** `src/soothsayer/oracle.py::Oracle.fair_value_lwc` (M6 deployed path); `Oracle.fair_value` (M5 reference path). Reads `data/processed/lwc_artefact_v1.parquet` (M6) or `data/processed/mondrian_artefact_v2.parquet` (M5) for the per-Friday `(symbol, fri_close, point, regime_pub, sigma_hat_sym_pre_fri)` tuple; the M6 16 scalars (M5 20 scalars) are hard-coded module constants.
2. **Rust serving crate.** `crates/soothsayer-oracle/src/oracle.rs::Oracle::fair_value` exposes both forecasters via `Oracle::load_with_forecaster(&path, Forecaster::{Mondrian, Lwc})`. The 16 LWC scalars and 20 Mondrian scalars are hard-coded in `config.rs`; the unit test `lwc_constants_match_sidecar` enforces ≤ 2 ULP agreement with `lwc_artefact_v1.json` on every refit.
3. **On-chain publisher.** `programs/soothsayer-oracle-program/`. The Rust crate is consumed by an Anchor program that emits `PriceUpdate` PDAs whose Borsh layout is preserved across the M5 → M6 migration. The wire `forecaster_code` byte distinguishes paths: code 2 = `FORECASTER_MONDRIAN` (M5; live on-chain); code 3 = `FORECASTER_LWC` (M6; live in the Rust publisher as of the Appendix E parity milestone, on-chain enablement gated on the next publisher release). The `soothsayer-consumer` no_std crate decodes the PDA into a typed `Forecaster::{Mondrian, Lwc}` view; consumers branching on `forecaster_code` need only add code-2 and code-3 cases.

`scripts/verify_rust_oracle.py` runs a dual-forecaster probe across 30 (symbol, fri_ts) cases × 3 target-coverage anchors per forecaster (90 cases per side, 180 total) and asserts byte-exact agreement on `(point, lower, upper, sharpness_bps, half_width_bps, claimed_served)` between Python and Rust on every case. **Submission status: 180/180 pass** (90/90 M5 + 90/90 M6) for the Python↔Rust parity. The Anchor program (`programs/soothsayer-oracle-program/tests/`) carries a scaffolded integration test that decodes the on-chain PDA after a publish call on the live M5 path; extending it to the M6 path is gated on the on-chain M6 enablement above.

A reviewer reproducing the paper end-to-end thus has three independent implementations of each forecaster that must agree on every (symbol, fri_ts, target_coverage) read.

A.9 Forward-tape harness operations

The forward-tape harness closes the $c(\tau)$ OOS-tuning loop on truly held-out data (the provenance disclosure is in §8, with detail in B.7): a content-addressed frozen artefact (`data/processed/lwc_artefact_v1_frozen_20260504.json`, SHA-256 `7b86d17a7691...`, freeze date 2026-05-04) is evaluated weekly against forward weekends as they accumulate. The frozen artefact's $\hat{\sigma}$ schedule, q_r table, and $c(\tau)$ are read-only; only the per-symbol $\hat{\sigma}_s(t)$ updates as new past-Fridays arrive.

Operations. The harness fires Tuesday mornings via launchd (`launchd/com.adamnoonan.soothsayer.forward-tape.plist`); each run executes a 26h-SLA pre-check on the upstream scryer feeds, runs `scripts/collect_forward_tape.py`, and writes `reports/m6_forward_tape_{N}weekends.md` with a "preliminary" banner when $N < 4$. The harness is silent about deployment; an out-of-band freeze re-roll is required to advance the artefact — after a planned methodology refresh, re-run `scripts/freeze_lwc_artefact.py` with a new `--date` and re-run the harness.

```
uv run python scripts/collect_forward_tape.py
uv run python scripts/run_forward_tape_evaluation.py
```

Status at submission (`reports/m6_forward_tape_11weekends.md`, generated 2026-07-14): $N = 11$ forward weekends, 2026-05-01 → 2026-07-10, $n = 110$ symbol-rows.

τ	n	realised	half-width (bps)	Kupiec p	Christoffersen p
0.68	110	0.7455	112.3	0.1330	0.7908
0.85	110	0.8455	179.2	0.8942	0.4758
0.95	110	0.9636	312.2	0.4912	0.3640
0.99	110	1.0000	515.6	0.1370	— (no violations)

Pooled Kupiec passes at all four anchors with no under-coverage; pooled Christoffersen passes at the three anchors with observed violations. Per-symbol Kupiec at $\tau = 0.95$ passes 10/10 on the forward tape, but at $n = 11$ per symbol the

test is powerless: two symbols (HOOD and MSTR) each realise a 18.2% violation rate (2 of 11), which Kupiec cannot reject at that n (the M5 in-sample baseline was 2/10). $N = 11$ clears the harness's own preliminary banner ($N \geq 4$) and remains just under the $N \geq 13$ threshold for cumulative pooled Christoffersen power. The submission's held-out backbone is §6.1's leave-one-symbol-out CV plus the nested temporal holdout, per-symbol Kupiec, split-date sensitivity, and four-year calendar sub-period stability (B.6); the forward-tape harness becomes load-bearing as N accumulates.

A sibling script (`scripts/run_forward_tape_variant_comparison.py`) carries an alternative-$\hat{\sigma}$ check on the same forward slice — never used to re-select, only to flag if a different $\hat{\sigma}$ rule looks dramatically cleaner on held-out data (§7, Appendix D).

B Full validation battery

This appendix carries the complete validation battery behind §6 and §8. The main text states the claims and the numbers that support them; this appendix walks the protocol, the full tables, and the diagnostics they are drawn from. Every table is emitted by a runner listed in A.3/A.5 and is reproducible from public data.

B.1 Evaluation protocol detail

The primary backtest panel comprises $N = 5,996$ weekend prediction windows: ten symbols × 639 weekends, 2014-01-17 through 2026-04-24 (the overnight panel of §6.4 / B.15 — $22,624$ rows on the same symbols — is constructed identically bar the gap selector; §5.2). Monday open is the target; Friday close and contemporaneous factor / vol-index readings compose $\mathcal{F}_t(s)$. The $\hat{\sigma}$ warm-up rule ($\geq$8 past observations per symbol) drops 80 rows at panel start to leave $5,916$ evaluable rows.

Four evaluation modes are reported: (i) an *in-sample machinery check* (quantile fit and served on the full evaluable panel; realised coverage matches τ by construction — B.2), (ii) the *out-of-sample validation* (quantile fit on $\mathcal{T}_{\text{train}} = \{t : t < \text{2023-01-01}\}$ — 4,186 rows; served on the 1,730-row $\mathcal{T}_{\text{test}}$ slice — B.3 onward), (iii) the held-out *forward-tape* evaluation against a content-addressed frozen artefact (§5.8, A.9), and (iv) a *simulation study* on synthetic panels with known ground truth (Appendix C). All p-values are two-sided Kupiec POF (proportion-of-failures) and Christoffersen independence tests. The base point estimator is unbiased (median residual -0.5 bps; MAE 90.0; RMSE 163.3); the product is the per-symbol-$\hat{\sigma}$-rescaled per-regime conformal quantile with the OOS-fit $c(\tau)$ bump.

B.2 In-sample machinery check

Full evaluable panel ($N = 5,916$):

Regime (n)	$\tau = 0.68$: realised / width (bps)	$\tau = 0.85$	$\tau = 0.95$	$\tau = 0.99$
normal (3,883)	0.685 / 91.0	0.851 / 142.7	0.951 / 247.7	0.987 / 424.8
long_weekend (619)	0.683 / 109.4	0.852 / 171.7	0.948 / 297.7	0.989 / 511.1
high_vol (1,414)	0.682 / 175.7	0.853 / 275.5	0.950 / 478.5	0.992 / 821.0
pooled (5,916)	**0.684 / 109.4**	**0.852 / 175.0**	**0.951 / 303.4**	**0.989 / 521.6**

Pooled realised coverage matches the target τ to within 0.2pp by construction. The narrower in-sample widths (vs the OOS table in B.3) reflect $\hat{\sigma}$'s tighter estimate over the full panel; the OOS table's wider bands reflect the OOS-fit $c(\tau) = 1.079$ at $\tau = 0.95$.

B.3 Out-of-sample per-regime coverage (2023+)

Quantile fit on pre-2023 weekends ($n = 4,186$); served on the 2023+ slice (1,730 rows × 173 weekends). The 12 trained per-regime quantiles $q_r(\tau)$ are held-out. The 4-scalar $c(\tau)$ schedule is OOS-fit, but the $\tau = 0.95$ headline is held out on two orthogonal axes — temporally via the nested TUNE/EVAL split and on symbol via leave-one-symbol-out CV (§6.1; detail in B.6) — both of which evaluate $c(0.95)$ on data its fit never touched. The walk-forward $\delta(\tau)$ schedule is identically zero (§4.5). B.7 carries the residual provenance disclosure.

Regime (n)	$\tau = 0.68$: realised / width	$\tau = 0.85$	$\tau = 0.95$	$\tau = 0.99$
normal (1,160)	0.687 / 109.7	0.847 / 172.3	0.947 / 300.1	0.990 / 471.6
long_weekend (190)	0.626 / 120.6	0.842 / 190.6	0.958 / 334.7	1.000 / 562.5
high_vol (380)	0.742 / 200.3	0.887 / 351.1	0.955 / 603.7	0.987 /1,169.9
pooled (1,730)	**0.693 / 130.8**	**0.855 / 213.6**	**0.950 / 370.6**	**0.990 / 635.0**

`high_vol` carries the widest bands, `normal` the narrowest, and every regime lands within ± 1pp of nominal at $\tau = 0.95$.

B.4 Conditional-coverage tests (pooled OOS)

τ	Violations	Rate	Kupiec p_{uc}	Christoffersen p_{ind}	$c(\tau)$
0.68	532	0.308	0.264	0.244	1.000
0.85	251	0.145	0.565	0.403	1.000
0.95	**86**	**0.050**	**0.956**	**0.603**	**1.079**
0.99	17	0.010	0.942	≈ 1.0	1.003

The $\tau = 0.95$ row is the headline pooled operating result. Realised coverage is exactly 0.9503 (Kupiec $p_{uc} = 0.956$); Christoffersen $p_{ind} = 0.603$ passes. Kupiec and Christoffersen pass at every served anchor on the deployed-fit OOS slice. Mean half-width at $\tau = 0.95$ is 370.6 bps — narrower than the K=26 $\hat{\sigma}$ comparator (385.3 bps; bootstrap 95% CI on Δhw% upper bound is -1.88%, see §7 / Appendix D). The held-out claim against this pooled baseline is the leave-one-symbol-out CV of §6.1 (0.9497 ± 0.0128). The residual Berkowitz / DQ rejection localises to per-symbol Berkowitz on TLT and TSLA (B.11) and to cross-sectional within-weekend common-mode (§8, B.8) — where the contract fails is §8's subject.

B.5 Realised-move tertile decomposition at $\tau = 0.95$

A post-hoc tertile labeler tags each weekend by realised-move z-score (calm / normal / shock); this `realized_bucket` is *not* a regime in the §3 sense — it depends on the realised target — and is used only for diagnostic stratification, including the shock-tertile ceiling reported in §8. Stratifying the OOS slice by realised-move $|z|$-score tertile (`reports/tables/m6_realised_move_tertile.csv`):

tertile (n)	realised	half-width (bps)	Kupiec p
calm (497)	0.9920	359.5	$< 10^{-6}$ (over-covers)
normal (601)	0.9933	379.6	$< 10^{-6}$ (over-covers)
shock (632)	**0.8766**	**370.7**	$< 10^{-6}$ (under-covers)
pooled (1,730)	**0.9503**	**370.6**	**0.956**

Half-width is approximately flat across tertiles (365–380 bps), so the band does not widen in shock periods — it just misses more often. The shock-tertile residual is cross-sectional common-mode within a weekend (§8), not per-symbol scale; pooled $\tau = 0.95 = 0.950$ is achieved through compensating tertile-level biases (calm 1.000, normal 0.993, shock 0.878). A $|z|$-conditional conformal upgrade (characterised in Appendix F) would tighten calm bands and widen shock bands at the same pooled τ.

B.6 Stability detail — nested holdout, TUNE anchors, split dates, walk-forward, calendar sub-periods

The headline stability results are stated in §6.1; this section carries the mode tables and mechanisms.

Nested temporal holdout — $c(\tau)$ fit on 2023, evaluated on 2024–26. TRAIN = pre-2023-01-01 (the unchanged 4,186-row per-regime quantile fit); TUNE = 2023-01-01 $\rightarrow$ 2023-12-29 (52 weekends, $c(\tau)$ fit slice); EVAL = 2024-01-05 $\rightarrow$ 2026-04-24 (1,210 rows × 121 weekends, true holdout). At $\tau = 0.95$:

mode	realised	$c(0.95)$	Kupiec p	Christoffersen p	hw bps	per-sym Kupiec
M_{full} (full-OOS c fit; B.4 baseline)	0.9504	1.079	0.956	0.720	370.6	10/10
M_{a2} (true temporal holdout)	**0.9504**	**1.079**	**0.947**	**0.989**	**420.8**	**10/10**

(The M_{full} row re-evaluates the B.4 deployed fit inside the nested-holdout runner; its Christoffersen statistic (0.720) differs from B.4's (0.603) because the lag-1 independence test is sensitive to the violation-series concatenation order — B.10 documents exactly this ordering sensitivity — and both pass at $\alpha = 0.05$. The reported Christoffersen figures use a fixed canonical ordering, weekend-major then symbol-lexicographic; the independence conclusion rests on no exact p-value, being corroborated order-robustly by the DQ and within-bin permutation tests, B.10 and B.16.) $c(\tau = 0.95)$ on TUNE-only matches the full-OOS fit at three decimals — the headline $\tau = 0.95$ number is *not* fit-on-evaluate sensitive. Christoffersen p improves in held-out mode (0.989 vs 0.720) because the 2024+ slice exhibits cleaner violation independence than the full 2023+ slice. The half-width difference (370.6 vs 420.8 bps) is entirely an eval-slice composition effect: the 2024+ holdout contains the 2024-08-05 BoJ unwind and the 2025-04 tariff weekend, both of which inflate $\hat{\sigma}_s$; coverage is unchanged because the $\hat{\sigma}_s$-conditional bands widened with the underlying realised volatility.

TUNE-anchor sensitivity. The result is stable across TUNE anchors $\{2021, 2022, 2023\}$: realised coverage at $\tau = 0.95$ lies in $[0.9474,\ 0.9524]$ with per-symbol Kupiec 10/10 preserved across all four anchors $\{2021, 2022, 2023, 2024\}$ (split-anchor robustness tables in Appendix D). The 2024 TUNE anchor over-covers at 0.9754 because TUNE-2024 includes the 2024-08-05 BoJ unwind, inflating $c(0.95)$ to 1.175; per-symbol Kupiec still passes 10/10. This is over-wide, not under-covering — the contract-favourable side of the asymmetry §8 frames.

Lower-τ over-coverage disclosure. At $\tau = 0.68$ and $\tau = 0.85$, M_{a2} realises 0.712 and 0.870 respectively (Kupiec $p = 0.015$ and $p = 0.044$, both rejecting toward over-coverage). Mechanism: $c(0.68)$ and $c(0.85)$ fit on TUNE-only inherit the elevated banking-crisis-weekend volatility of 2023 (March SVB/CS), inflating c to 1.028 and 1.026 respectively; M_{full}'s schedule on the full OOS slice averages this out at the floor ($c = 1.000$) and is the deployed mitigation. Source: `reports/tables/paper1_a2_nested_holdout_c_tau.csv`, `paper1_a2_6_tune_anchor_robustness.csv`.

Split-date sensitivity (stability-of-fit on the deployed slice). Repeating the fit (quantile table re-trained, $c(\tau)$ re-fit per split, $\delta \equiv 0$ held) at four OOS-split anchors $\{2021\text{-}01\text{-}01, 2022\text{-}01\text{-}01, 2023\text{-}01\text{-}01, 2024\text{-}01\text{-}01\}$ delivers realised $\tau = 0.95$ coverage of $\{0.9539,\ 0.9520,\ 0.9503,\ 0.9504\}$ — Kupiec $p \in \{0.348,\ 0.661,\ 0.956,\ 0.947\}$ and Christoffersen $p \in \{0.115,\ 0.186,\ 0.603,\ 0.671\}$ — every (split × τ) cell passes at $\alpha = 0.05$. Mean half-width $\{357.2,\ 363.5,\ 370.6,\ 416.8\}$ bps varies $\pm 8\%$ around the deployed value.

Walk-forward stability. Re-running the schedule selection on the four powered expanding-window splits (fractions $f \in \{0.40, 0.50, 0.60, 0.70\}$ — the 0.20 and 0.30 splits are excluded as under-powered for the 4-scalar $c(\tau)$ fit; Appendix D) with $\delta = 0$: walk-forward realised $\tau = 0.95$ coverage is ≥ 0.95 on every powered split (range 0.954–0.975) — the criterion that selected $\delta = 0$ (§4.5). Per-split mean half-width at $\tau = 0.95$ tracks the full-OOS-fit value within $\pm 8\%$.

Calendar sub-period stability. Bucketing the OOS slice into four calendar sub-periods {2023, 2024, 2025, 2026-YTD} (`reports/tables/m6_subperiod_robustness.csv`):

Sub-period (n weekends)	Realised at $\tau = 0.95$	Half-width (bps)	Kupiec p	Christoffersen p
2023 (52)	0.950	253.6	1.000	0.888
2024 (52)	0.931	420.4	0.057	0.841
2025 (52)	0.971	444.4	0.017	0.793
2026-YTD (17)	0.947	350.0	0.862	0.908
Pooled OOS (173)	**0.950**	**370.6**	**0.956**	**0.603**

Pooled OOS calibration at $\tau = 0.95$ holds across each of the four years that span SVB / banking-stress 2023, the 2024 rate-cut transition, the 2025 tokenisation-launch year, and 2026-YTD (realised range 0.931–0.971). The 2025 cell rejects Kupiec at $\alpha = 0.05$ ($p = 0.017$, over-coverage); 2024 sits just above the threshold ($p = 0.057$); 2023 and 2026-YTD pass cleanly. Across the full 16-cell (sub-period × τ) grid the architecture carries 3 Kupiec rejections — all toward over-coverage (one 2024 cell at $\tau = 0.99$ and the 2025 cells at $\tau \in \{0.95, 0.99\}$; §8 frames the asymmetry); Christoffersen lag-1 independence is not rejected in any of the 16 cells.

Stationarity read. The P_2 guarantee (§3) assumes the standardised score's conditional distribution is approximately stationary; the sub-period grid is the direct empirical evidence, and it is positive: four years of structurally different market regimes, three rejections, all on the over-coverage side — not the under-coverage failure mode that warrants a deployment alarm. The EWMA $\hat{\sigma}_s(t)$ provides partial adaptive scaling per symbol (the Appendix C simulation shows $\hat{\sigma}$ tracks drift and structural-break DGPs with a half-life-bounded lag), but conditional-distribution stationarity of the *standardised* score is still assumed; §8 lists the three violations no backtest grid can pre-empt.

B.7 OOS-tuned $c(\tau)$ provenance

Of the four OOS-fit $c(\tau)$ scalars, three are essentially identity ($c \in \{1.000,\ 1.000,\ 1.003\}$ at $\tau \in \{0.68,\ 0.85,\ 0.99\}$); **only** $c(0.95) = 1.079$ **carries meaningful OOS information** (a 7.9% widening at the headline anchor). The nested temporal holdout (B.6) fits $c(0.95)$ on 2023 alone and evaluates it on a 2024-01-05 $\rightarrow$ 2026-04-24 slice that neither the per-regime quantile nor the bump has seen — realised coverage 0.9504, Kupiec $p = 0.947$, Christoffersen $p = 0.989$, per-symbol Kupiec 10/10; $c(0.95)$ on TUNE-only matches the full-OOS fit at three decimals. **The headline $c(0.95) = 1.079$ is empirically held-out at the time level.** The residual provenance disclosure is restricted to the lower-τ over-coverage in held-out mode (B.6), which the deployed full-OOS schedule mitigates at the floor. The forward-tape harness (A.9) closes the residual contamination by carrying the frozen deployment artefact onto truly held-out weekends as they accumulate: at submission $N = 11$ forward weekends show pooled Kupiec passing at all four anchors with no under-coverage and per-symbol Kupiec 10/10 at $\tau = 0.95$ — just under the $N \geq 13$ gate at which the disclosure upgrades to "OOS-fit + held-out forward-tape re-validation."

B.8 Joint-tail empirical distribution — fitting detail and stability

§8 states the joint-tail result and carries Figure H5; this section documents the baseline construction and the full statistics. For each OOS weekend w with full 10-symbol coverage, $k_w = |\{s : s \text{ breaches its } \tau\text{-band on } w\}|$ is compared against three properly specified baselines: $\mathrm{Binom}(10, 1-\tau)$ (independence), a Gaussian copula on the empirical correlation matrix $\hat{R}$ (mean off-diagonal 0.36 on standardised residuals), and a Student-t copula on per-symbol Student-t marginals with $\hat{\nu} \in [4.80, 28.32]$, copula degrees-of-freedom $\hat{\nu}_{\text{copula}} = 6.04$ (median across symbols) (`reports/tables/paper1_a3_joint_baseline_kw_distribution.csv`, `..._summary.csv`; 50,000 simulated weekends per baseline; CIs are 1000-rep weekend-block bootstrap, seed = 0):

Statistic at $\tau = 0.95$ ($n_{\text{weekends}} = 173$)	empirical	$\mathrm{Binom}(10, 0.05)$	Gaussian copula	t-copula ($\nu = 6.04$)
Mean k_w / weekend	0.497	0.500	0.501	0.501
Variance ratio (emp / model)	—	**2.34**	0.83	**0.63**
$P(k_w \geq 3)$	**4.62% [1.73, 7.51]**	1.15%	6.69%	8.15%
$P(k_w \geq 5)$	**0.58% [0.00, 1.73]**	0.01%	1.86%	3.15%
max k_w (worst weekend)	9	—	—	—

Variance overdispersion. Empirical k_w over-disperses by a uniform $\sim 2.3\times$ vs Binom independence at every served τ (emp/Binom variance ratio $\{2.34, 2.31, 2.34, 2.35\}$ at $\tau \in \{0.68, 0.85, 0.95, 0.99\}$) and sits inside a properly-specified t-copula envelope from above (emp/t variance ratio $\{0.83, 0.54, 0.63, 0.73\}$; Wasserstein-1 distance to the t-copula ≤ 0.180 at every τ, less than half the distance to Binom independence). Marginal calibration is preserved — mean k_w matches every reference to 0.003. Independence is misspecified, the joint dependence is real ($\hat{R} = 0.36$ mean off-diagonal), and the structure brackets-bound: empirical between Binom independence and a homogeneous t-copula

at $\hat{\nu} \approx 6$. The mean intra-weekend pairwise breach correlation ($\bar{\rho}_{\text{intra}} = 0.41$ — the average pairwise dependence across the ten symbols within a weekend, distinct from the lag-1 cross-sectional AR coefficient $\hat{\rho}_{\text{cross}} = 0.354$ of B.10) is respected by all reported CIs via weekend-block bootstrap; temporal autocorrelation across weekends is empirically null (lag-1, lag-2, lag-4 $\hat{\rho}$ on weekly k_w in $[-0.07, 0.03]$), so $L \in \{4, 8, 13\}$ moving-block-bootstrap CIs are within Monte Carlo noise of the $L = 1$ baseline (`reports/tables/paper1_b5_kw_block_bootstrap.csv`).

Reserve-guidance threshold stability. k^* at each τ — smallest k whose empirical hit-rate is closest to 5% — is $k^* = 5$ at $\tau = 0.85$ (rate 5.78%), $k^* = 3$ at $\tau = 0.95$ (4.62%), and $k^* = 1$ at $\tau = 0.99$ (6.36%). Across the 24-cell grid (3 τ × 2 threshold conventions × 4 sub-periods) the architecture carries **0/24 Kupiec stability rejections** at $\alpha = 0.05$ (`reports/tables/m6_kw_threshold_stability.csv`); at $k^* = 3$ for $\tau = 0.95$, per-sub-period hit-rates are $\{3.85, 7.69, 1.92, 5.88\}\%$ across {2023, 2024, 2025, 2026-YTD} with Kupiec $p \in \{0.78, 0.33, 0.30, 0.81\}$. The per-sub-period 95th percentile of k_w at $\tau = 0.95$ stays in $[1.0, 3.0]$ across the four years. With sub-period $n \approx 52$ and target rates $\sim 5\%$, the binomial-stability test has limited power against drift below the year-on-year p95 spread (≤ 1.6 integer-symbols); the result is "no detectable instability with the available power." At $\tau = 0.85$ the empirical 99th percentile of k_w is ≈ 7 vs binomial ≈ 4; §8 carries the consumer guidance.

B.9 Worst-observed weekend — 2024-08-05 BoJ yen-carry unwind

The single OOS weekend with all 10 symbols breaching the $\tau = 0.85$ band simultaneously is **Friday 2024-08-02 → Monday 2024-08-05** — the Bank of Japan rate-hike yen-carry-trade unwind, one of the largest single-weekend global risk-off events of the post-2020 period. The same weekend has $k_w = 9$ at $\tau = 0.95$ (only TLT inside that band — consistent with the B.8 worst-weekend maximum) and $k_w = 5$ at $\tau = 0.99$. §8 carries the one-paragraph summary; the per-symbol anatomy follows. The deployed served band, computed directly from the deployment artefact (per-Friday parquet row + sidecar schedules, EWMA HL=8 $\hat{\sigma}$) on the 2024-08-02 row:

Symbol	Weekend return	HW @ $\tau = 0.85$	Breach @ $\tau = 0.85$	@ $\tau = 0.95$
MSTR	-27.4%	663 bps	-1537 bps	breach (-1060)
HOOD	-17.8%	357 bps	-1336 bps	breach (-1079)
NVDA	-14.2%	332 bps	-1000 bps	breach (-762)
TSLA	-10.8%	499 bps	-496 bps	breach (-137)
AAPL	-9.4%	201 bps	-659 bps	breach (-514)
GOOGL	-6.7%	194 bps	-390 bps	breach (-250)
QQQ	-5.4%	120 bps	-330 bps	breach (-243)
SPY	-4.0%	85 bps	-228 bps	breach (-167)
GLD	-2.1%	108 bps	-171 bps	breach (-94)
TLT	$+1.4\%$	131 bps	$+10$ (just past)	inside

Cross-section: 9 of 10 symbols sold off in concert (mean weekend return -963 bps, median -807 bps); only TLT delivered the classic flight-to-quality bid. Macro context: the 2024-08-02 US July nonfarm-payrolls print landed at 114k vs ~175k consensus with the Sahm-rule recession indicator triggering; the 2024-07-31 BoJ rate hike combined with hawkish Powell/Yellen language drove an accelerating yen-carry unwind through Asian time zones over the weekend; the Monday 2024-08-05 Nikkei fell 12.4% (largest single-day drop since Black Monday 1987) and intraday VIX spiked to ~65, the highest reading since the COVID-19 March 2020 panic.

The $\hat{\sigma}$-standardisation held at the per-symbol level — high-vol symbols (MSTR, TSLA) got wide bands, low-vol symbols (SPY, QQQ, GLD) narrow ones, and breach magnitudes scale with the cross-sectional shock divided by $\hat{\sigma}$. What broke is *cross-sectional common-mode*: when nearly every symbol moves the same direction at the same time by 4–27 standard-deviation-scale weekend returns, no per-symbol band — however well calibrated — can absorb the joint event. A consumer monitoring k_w in real time would have seen $k_w = 10$ at $\tau = 0.85$ by the Monday-open print, well in excess of the deployment-time-fitted $k^* = 5$ (or the empirical 99th percentile of 7); the empirical k_w distribution is the right operational signal precisely because it captures the cross-sectional common-mode that per-symbol coverage cannot.

Joint-baseline framing of the BoJ probability. Under a Student-t copula with $\hat{\nu} = 6.04$ and empirical correlation $\hat{R}$ (mean off-diagonal 0.36), $P(k_w = 10)$ at $\tau = 0.85$ is 0.21% — i.e., the 2024-08-05 BoJ unwind is a $\sim$ 1-in-475-weekends event under a properly specified joint baseline. The independence baseline ($\text{Binom}(10, 0.15)$) puts the same all-ten-breach event at $0.15^{10} \approx 5.8 \times 10^{-9}$, i.e. $\sim$ 1-in-1.7×10^8-weekends; the residual structure §8 documents is exactly the gap between these two figures. The residual has a *name* (cluster topology, §8), a *parametric*

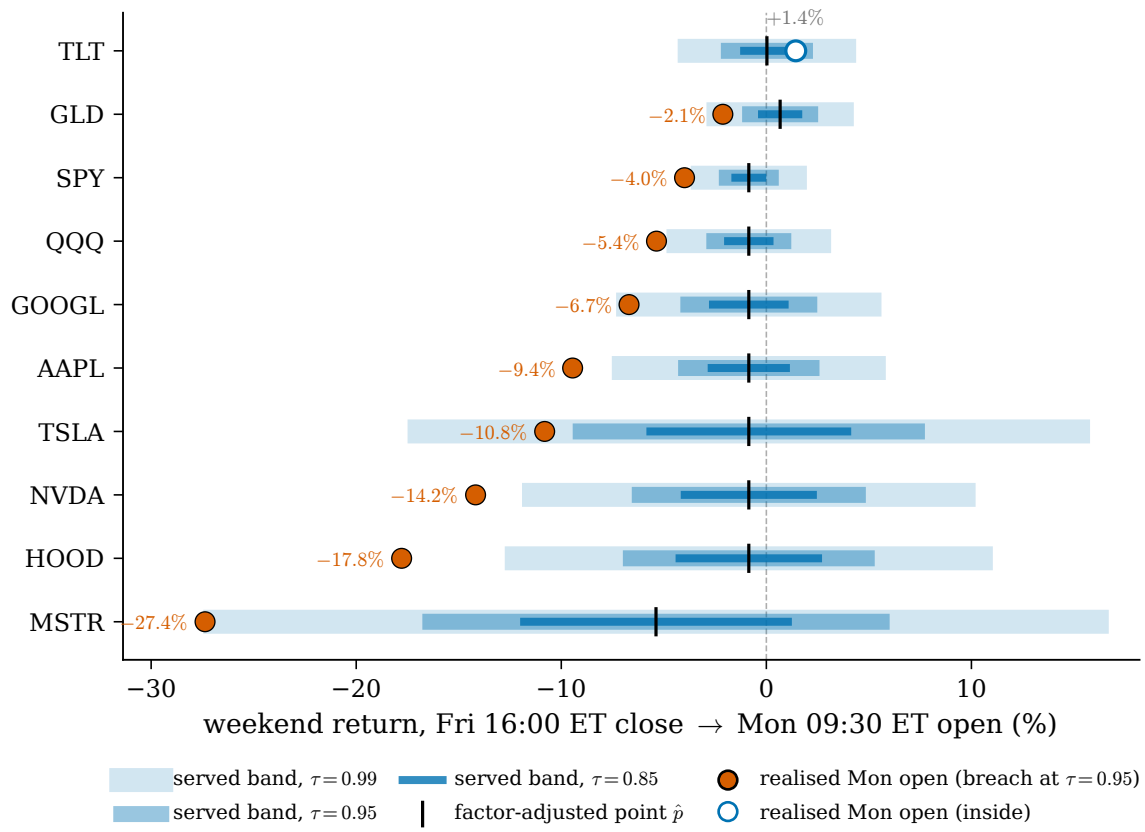

Figure 10: Anatomy of the served band on the worst observed weekend (2024-08-02 → 2024-08-05, BoJ yen-carry unwind). For each symbol: nested served bands at $\tau \in \{0.85, 0.95, 0.99\}$ (blue, darkest = 0.85), the factor-adjusted point $\hat{p}$ (black tick), and the realised Monday open (filled vermilion = breach at $\tau = 0.95$; open = inside), in weekend-return space relative to Friday close. Bands are computed from the deployment artefact, byte-aligned with what a consumer read. Per-symbol $\hat{\sigma}_s$ width differentiation is visible directly — MSTR's $\tau = 0.85$ half-width (663 bps) is $\approx 8\times$ SPY's (85 bps) — and so is the cross-sectional common-mode failure: nine realised opens march left past their bands in concert, which no per-symbol band, however well calibrated, can absorb. The k_w distribution of §8 / B.8 is the operational handle for exactly this event class.

envelope (t-copula at $\nu \approx 6$ with empirical $\hat{R}$), and a *structural mechanism* (equity vs safe-haven cluster topology) — not "a residual the architecture cannot characterise."

B.9.1 Case study — the M6 served bands against hypothetical consumer configurations

A per-method coverage grid on the same weekend makes the §8 claim concrete: no method on the grid attains its nominal coverage at $\tau \leq 0.95$ on this weekend. Generated by the self-checking runner `scripts/build_case_study_boj_m6.py` (band logic mirrors `build_paper1_figures.py::fig9_boj_anatomy`; sanity-gated against B.9's $k_w = 10/9/5$; input hashes recorded in the report header). Bands are read from the deployment artefact, not refit — byte-aligned with what a consumer read on 2024-08-02; the 2024-08-02 row sits in the OOS slice.

The comparators are **hypothetical consumer configurations, not incumbent products**: "Pyth+k%" wraps a Pyth-style point price in a fixed symmetric k% buffer; "VIX-scaled" and "Const-buffer" are this project's v1-era calibrated baselines (pre-2023 train, centred on Friday close), whose half-widths are retained for comparability. Coverage = Monday 09:30 ET open inside the band. Regime classification for 2024-08-02 = `high_vol` for all 10 symbols. Realised moves: max |Mon−Fri|/Fri = 2,737 bps (MSTR); mean |move| = 992 bps; 7 of 10 symbols exceed 500 bps.

The per-symbol deployed-band (M6) view — coverage ✓/✗ and half-width (bps) at each served τ; the full grid including every comparator column is in `case_study_boj_m6.md`:

Symbol	Realised (bps)	M6 τ=0.68	M6 τ=0.85	M6 τ=0.95	M6 τ=0.99
AAPL	−945	✗ / 114	✗ / 201	✗ / 345	✗ / 668
GLD	−212	✗ / 62	✗ / 108	✗ / 185	✓ / 359
GOOGL	−670	✗ / 111	✗ / 194	✗ / 334	✓ / 647
HOOD	−1779	✗ / 204	✗ / 357	✗ / 614	✗ / 1190
MSTR	−2737	✗ / 378	✗ / 663	✗ / 1139	✓ / 2208
NVDA	−1418	✗ / 189	✗ / 332	✗ / 571	✗ / 1106
QQQ	−536	✗ / 69	✗ / 120	✗ / 207	✗ / 402
SPY	−399	✗ / 49	✗ / 85	✗ / 146	✗ / 283
TLT	+143	✗ / 75	✗ / 131	✓ / 225	✓ / 435
TSLA	−1081	✗ / 285	✗ / 499	✗ / 859	✓ / 1664

Aggregate coverage on this weekend across all methods (✓ count / total):

Method	Covered	Mean half-width (bps)
Pyth+2%	1/10	200
Pyth+5%	3/10	500
Pyth+10%	6/10	1000
Pyth+20%	9/10	2000
VIX-scaled τ=0.68	0/10	103
VIX-scaled τ=0.85	1/10	173
VIX-scaled τ=0.95	2/10	329
VIX-scaled τ=0.99	5/10	700
Const-buffer τ=0.68	0/10	75
Const-buffer τ=0.85	0/10	139
Const-buffer τ=0.95	2/10	272
Const-buffer τ=0.99	5/10	694
M6 τ=0.68	0/10	153
M6 τ=0.85	0/10	269
M6 τ=0.95	1/10	463
M6 τ=0.99	5/10	896

Reading the grid:

- **M6 at $\tau = 0.68$ and $\tau = 0.85$ covers 0/10.** All ten symbols breach the $\tau = 0.85$ band simultaneously; this weekend is the single $k_w = 10$ event at $\tau = 0.85$ in the OOS record (§8, B.8).
- **M6 at $\tau = 0.95$ covers 1/10** (TLT only; $k_w = 9$, the OOS maximum at that anchor). The $\hat{\sigma}$ standardisation works at the per-symbol level — MSTR's $\tau = 0.85$ half-width (663 bps) is $\approx 8\times$ SPY's (85 bps) — but nine symbols moving the same direction by 4–27 $\hat{\sigma}$-scale weekend returns is a cross-sectional common-mode event that no per-symbol band absorbs. The joint-breach distribution ($k^* = 3$ reserve-guidance threshold at $\tau = 0.95$; §8) is the operational handle for this event class.
- **M6 at $\tau = 0.99$ covers 5/10** (GLD, GOOGL, MSTR, TLT, TSLA; mean half-width 896 bps). Coverage at this anchor comes from per-symbol width differentiation plus the factor-adjusted centre: MSTR's served band is 2,208 bps wide and centred −538 bps below Friday close, so its −2,737 bps realised move lands inside; SPY's 283 bps band does not reach its −399 bps move.
- **Fixed-buffer comparators:** Pyth+5% covers 3/10; Pyth+10% covers 6/10 and Pyth+20% covers 9/10, at uniform 1,000 / 2,000 bps half-widths on every symbol in every week regardless of regime. The v1-era calibrated comparators (VIX-scaled, const-buffer; pre-2023 train) cover at most 2/10 at $\tau \leq 0.95$ and 5/10 at $\tau = 0.99$.
- **No method on this grid attains its nominal coverage at $\tau \leq 0.95$ on this weekend.** The best $\tau \leq 0.95$ cell is 2/10. The tail-coverage story on this event sits at $\tau = 0.99$, where M6 matches the best comparator count (5/10) with regime- and symbol-conditional widths rather than a flat buffer.

One weekend is one observation. The aggregate OOS calibration evidence for the deployed architecture is §6 and B.3–B.8; this case study is the qualitative counterpart on the worst observed weekend.

B.10 Extended diagnostics — DQ and Berkowitz specification

Engle-Manganelli (2004) DQ is the finer test on the violation series. Kupiec checks marginal counts and Christoffersen's two-state Markov independence checks the lag-1 hit transition; neither has power against conditional miscalibration patterns that yield correct marginal counts and uncorrelated lag-1 hits but encode systematic dependence on lagged hits or the served quantile [Engle and Manganelli, 2004]. DQ regresses the centred hit indicator on a constant + K lagged hits and tests joint significance under χ^2_{K+1}. We use $K = 4$ throughout (the lag-only specification of Engle–Manganelli's CAViaR — conditional autoregressive value-at-risk — Table 1, block 1), no contemporaneous covariate — the lag-only specification is symmetric across the five methods compared in §6.2, so cross-method DQ p-values are directly comparable. **Pooled DQ at $\tau = 0.95$ rejects on the deployed fit** — the cross-sectional within-weekend common-mode (orthogonal to $\hat{\sigma}$ standardisation) is what DQ is sensitive to here. Per-(symbol × τ × method) DQ p-values are reported in the §6.2 master grid alongside Kupiec and Christoffersen. **Bands are *per-anchor calibrated*, not full-distribution calibrated** — the joint-tail empirical distribution (§8, B.8) turns this disclosure into a measurable, reportable shape.

Berkowitz (2001) joint LR on continuous PITs. PITs reconstructed by inverting each served band against the realised Monday open through the per-regime conformal CDF. The pooled Berkowitz LR on the deployed PITs is non-zero, viewing the same residual through the PIT lens. The Berkowitz/DQ rejection and the B.8 joint-tail overdispersion are not two independent residuals: both are the cross-sectional within-weekend common-mode ($\hat{\rho}_{\text{cross}} = 0.354$) viewed through different aggregations — Berkowitz/DQ on the pooled PIT/violation series detects it as serial-cluster-equivalent dependence; the k_w distribution on the same panel detects it as upper-tail overdispersion in joint-breach counts. The methodology fix targets the same channel; the consumer-facing handle is the reserve-guidance threshold (§8).

Localising the residual. Decomposing the lag-1 alternative under panel-row order: the AR(1) signal lives in the *cross-sectional within-weekend* ordering ($\hat{\rho}_{\text{cross}} = 0.354$, $p < 10^{-100}$, $n = 1,557$) — not the temporal-within-symbol ordering ($\hat{\rho}_{\text{time}} = -0.032$, $p = 0.18$, $n = 1,720$). $\hat{\sigma}$ standardisation operates *within* a symbol's history, not *across* symbols within a weekend, so it cannot reach this residual. A vol-tertile sub-split of `normal` regime (5 cells) leaves Berkowitz LR essentially unchanged ($170 \to 172$) while widening the $\tau = 0.95$ band by 9% — finer regime granularity is not the lever (`reports/tables/m6_lwc_robustness_vol_tertile.csv`). The candidate architectures that target cross-sectional common-mode (cluster-conditional partial-out, Appendix F) are deferred-with-gates.

B.11 Per-symbol Kupiec at $\tau = 0.95$, Berkowitz LR, and test power

Per-symbol Kupiec on the 2023+ OOS slice (`reports/tables/m6_per_symbol_kupiec_4methods.csv`, regenerated 2026-05-05):

Symbol	Violation rate	Kupiec p	Berkowitz LR
AAPL	0.069	0.268	6.7
GLD	0.069	0.268	3.2
GOOGL	0.052	0.903	10.2
HOOD	0.035	0.329	6.4
MSTR	0.046	0.818	7.1
NVDA	0.052	0.903	10.0
QQQ	0.040	0.552	5.1
SPY	0.046	0.818	9.6
TLT	0.046	0.818	16.7
TSLA	0.040	0.552	13.5
pass-rate at $\alpha = 0.05$		**10 / 10**	
Berkowitz LR range			**3.2 – 16.7** ($5.2\times$)

All 10 symbols pass Kupiec at $\tau = 0.95$ with violation rates inside $[3.5\%, 6.9\%]$. Within the per-symbol Berkowitz LR range $[3.2, 16.7]$ (a $5.2\times$ spread), only TLT (LR 16.7) and TSLA (LR 13.5) reject Berkowitz at $\alpha = 0.01$, and those rejections trace to cross-sectional common-mode rather than per-symbol scale (B.10).

Power caveat at $\tau = 0.99$. Per-symbol Kupiec at $\tau = 0.99$ has minimum detectable effect ≈ 3 pp under-coverage and is undetectable for over-coverage given the per-symbol expected violation count of 1.73 (Monte Carlo: 10,000 reps per cell, Kupiec LR critical at $\alpha = 0.05$; `reports/tables/paper1_b3_kupiec_power_mde.csv`). The 10/10 pass should be read as "no per-symbol violation rate is statistically distinguishable from 1% at the resolution Kupiec offers per-symbol", **not** as a 1 pp tolerance certificate. Pooled Kupiec at $\tau = 0.99$ has minimum detectable effect 1.0

pp; the pooled $p = 0.94$ is the $\tau = 0.99$ statistical claim that does have 1 pp resolution. Per-symbol Kupiec at $\tau = 0.95$ has MDE 4.0 pp under-coverage / 7.5 pp over-coverage on $N = 173$ — the deployed per-symbol violation rates in $[3.5\%, 6.9\%]$ sit well within the 4 pp window of nominal 5%, so the 10/10 pass at $\tau = 0.95$ is informative against deviations larger than 4–8 pp.

B.12 GARCH baselines — pooled tables, matched-coverage width, proper scoring

The textbook econometric default for a time-varying interval at τ is per-symbol GARCH(1,1) on log Friday-to-Monday returns. The practitioner choice of innovation distribution is Student-t: Gaussian innovations under-fit equity weekend tails, where the fitted $\hat{\nu} \approx 3$ across the panel (`reports/tables/m6_lwc_robustness_garch_t_baseline.csv`):

τ	Method	Realised	HW (bps)	Kupiec p	Christoffersen p
0.95	GARCH-Gaussian	0.9254	322.2	0.000	0.016
0.95	GARCH-t	0.9277	331.6	0.000	**0.209**
0.95	**this paper**	**0.9503**	**370.6**	**0.956**	**0.603**
0.99	GARCH-Gaussian	0.9630	423.7	0.000	0.866
0.99	GARCH-t	0.9850	569.3	0.050	1.000
0.99	**this paper**	**0.9902**	**635.0**	**0.942**	1.000

GARCH-t closes two failures GARCH-Gaussian leaves on the table: the $\tau = 0.99$ tail capture (realised improves from 0.963 to 0.985; Kupiec p moves from < 0.001 to a borderline 0.050) and the Christoffersen rejection at $\tau = 0.95$ (p moves from 0.016 to 0.209). What it does *not* fix is the **Kupiec under-coverage at every** $\tau < 0.99$ — the standardised-t 95th percentile at $\nu \approx 3$ is ~ 1.83 vs Gaussian 1.96, so the bands tighten precisely where Gaussian was already under-covering. T-innovations fix the wrong end of the distribution.

At matched 95% realised coverage, widening GARCH-t to its claimed $\tau = 0.95$ requires roughly $+14\%$ on width, giving a matched-coverage HW ≈ 378 bps — comparable to the deployed architecture's 370.6 bps (we report the two half-widths rather than a paired width test; the claim rests on the coverage difference, not a width difference). The distinction that matters is provenance: the deployed 370.6 bps is served with its coverage stated in advance, while GARCH-t's ≈ 378 bps is the width that reproduces the target coverage only in hindsight. At $\tau = 0.99$ the deployed architecture dominates GARCH-t on coverage and width-at-matched-coverage simultaneously. The conformal route fixes both Kupiec and Christoffersen because it is calibrated to the data, not to a fitted parametric distribution; the NVDA fallback (t-MLE pushes $\hat{\nu}$ to the optimizer lower bound $\hat{\nu} = 2.50$, adjacent to the $\nu \leq 2$ variance-undefined region) is the parametric-tail risk the deployed architecture avoids.

Proper-scoring-rule head-to-head. Under proper scoring rules — Winkler (1972) interval score and CRPS (continuous ranked probability score) — the deployed architecture dominates GARCH(1,1)-t at every served τ. Winkler interval score (bps of fri_close, lower is better) at $\tau = 0.95$: deployed 992 vs GARCH-t $1,139$ (-12.9%); at $\tau = 0.99$: deployed $1,566$ vs GARCH-t $1,904$ (-17.8%). The advantage grows with τ — the $\hat{\sigma}$-standardisation pays most where parametric-tail mis-specification hurts most. Pooled CRPS over the served coverage range $\{0.05, 0.10, \ldots, 0.99\}$: deployed 80.7 vs GARCH-t 92.6 (-12.8%). Both rules are computed on the aligned 1,730-row OOS slice; method aggregation is by row, not by τ, so a single number summarises the practitioner-baseline comparison without anchor-row cherry-picking. Source: `reports/tables/paper1_c1_winkler_interval_score.csv`, `..._crps.csv`. The Winkler comparison against the *tokenised-tracking* baseline archetype — the competitor motivated by Cong et al.'s conditional-mean result, distinct from the GARCH-family parametric baselines compared here — is carried in §6.3 with construction detail in Appendix D.

B.13 Per-asset-class deviation

Pooled OOS coverage stratified by asset class under the deployed EWMA HL=8 $\hat{\sigma}$ (`reports/tables/m6_lwc_robustness_per_class.csv`). Equities (8 syms, 1,384 obs): realised 0.952 at $\tau = 0.95$ with hw 417.8 bps; GLD ($n = 173$): realised 0.931, hw 180.9 bps; TLT ($n = 173$): realised 0.954, hw 182.3 bps. Row-weighted reconciliation: $(1384 \cdot 417.8 + 173 \cdot 180.9 + 173 \cdot 182.3)/1730 = 370.6$ bps, matching the B.4 panel-pooled HW. The per-symbol $\hat{\sigma}_s$ multiplier delivers heavy-tail equities wider bands matched to their relative-residual scale and the defensive class (GLD, TLT) narrower bands matched to theirs; the deployed architecture's Kupiec passes every class at every τ. The width allocation across asset classes is the operational manifestation of P_3 (per-regime serving efficiency): width is concentrated where calibration demands it.

B.14 Path coverage — endpoint vs intra-weekend

The §6 result is *endpoint coverage*: realised Monday open lies inside the served band. A DeFi consumer holding an xStock as collateral is exposed at every block over the weekend; a Saturday liquidation when the on-chain xStock briefly trades outside the band is a real loss event even if Monday open returns inside. We measure path coverage on [Fri 16:00 ET, Mon 09:30 ET] against 24/7 stock-perp 1m bars from `cex_stock_perp/ohlcv` (Kraken Futures `PF_<sym>XUSD`, xStock-backed). Sample: 19 weekends × 9 symbols (TLT excluded — no perp listing), $n = 118$ (symbol, weekend) pairs per anchor, 2025-12-19 → 2026-04-24 (`reports/tables/path_coverage_perp.csv`):

τ	n	Endpoint coverage	Path coverage	Gap (pp)
0.68	118	0.644	0.348	+29.7
0.85	118	0.864	0.568	+29.7
0.95	**118**	**0.949**	**0.788**	**+16.1**
0.99	118	0.992	0.915	+7.6

The deployed architecture calibrates the perp-listed sample to nominal endpoint coverage (0.949 at $\tau = 0.95$ vs nominal 0.95) on a sample whose composition is predominantly large-cap equities (SPY/QQQ/AAPL/GOOGL/NVDA/MSTR/TSLA/HOOD plus GLD; TLT excluded — no perp listing). The path-coverage gap is $+16.1$ pp at $\tau = 0.95$: an honestly-calibrated endpoint band exhibits a structural shortfall on a sample whose intra-weekend variance is larger than its endpoint variance. Sample is small (binomial CI on the $\tau = 0.95$ pooled gap is ± 6 pp); the test is directional. After three confound checks documented in `reports/v1b_path_coverage_robustness.md` — perp-spot basis normalisation closes the gap to 11.0 pp, a volume-floor filter at ≥ 1 contract closes it to 9.5 pp ($n = 63$), and a 15-minute rolling-median sustained-crossing definition leaves it at 14.4 pp — the residual genuine-shortfall band is $\sim$ 9–15 pp at $\tau = 0.95$, with most of the magnitude attributable to sustained drift rather than transient thin-liquidity prints. The gap collapses meaningfully only at $\tau = 0.99$ (+7.6 pp); continued capture is tracked under scryer item 51, and the path-fitted conformity score is the methodology variant (Appendix F).

Decomposing the gap: fair-value path vs venue microstructure. An excursion-inflation diagnostic separates the two sources of the perp gap. Define $\lambda(\tau)$ as the multiplier on the endpoint band half-width required to reach τ *path* coverage. On the CME factor-projected path — the projectable weekend trajectory of the per-symbol factor scaled to underlier units, 435 weekends — $\lambda(\tau) \approx 1.0$ at every anchor: the endpoint-sized band already contains the factor's intra-weekend round-trips (`reports/active/proto_excursion_inflation.md`). The larger perp λ ($\approx$ 1.6–2.4, and non-monotone in τ at this n) therefore reflects predominantly *venue basis and microstructure* on a thin-liquidity perp — not fair-value excursion the forecaster fails to capture, consistent with the thin-print confounds above. The contract implication: the deployed band is an **endpoint** contract that, on the projectable fair-value path, also delivers $\approx \tau$ path coverage; the residual a continuous-consumption consumer experiences is a property of their settlement venue, and closing it on a real venue is a microstructure layer (the oracle-conditioned-AMM setting), not a forecaster deficiency.

The endpoint contract stands. A consumer requiring path coverage at level τ should step up one anchor (empirically closes the gap on this slice at $\tau = 0.99$), absorb the residual through their own downstream risk policy, or — for continuous-consumption AMM use cases that cannot use the step-up lever — adopt the path-fitted conformity-score variant (Appendix F).

B.15 Overnight generalisation — full battery

§6.4 states the overnight result; this section carries the tables and the cadence-specific data treatments. Pooled OOS conditional coverage (overnight; TRAIN < 2023-01-01, 16,093 rows; OOS $\geq$ 2023-01-01, 6,450 rows × 645 nights; post-dividend-adjustment run, `reports/active/overnight_calibration_firstread.md` @ 2026-06-25):

τ	Violations	Rate	Kupiec p_{uc}	Christoffersen p_{ind}	$c(\tau)$
0.68	2,064	0.320	1.000	0.537	1.019
0.85	963	0.149	0.875	0.212	1.000
0.95	**316**	**0.049**	**0.710**	**0.158**	**1.000**
0.99	53	0.008	0.138	1.000	1.000

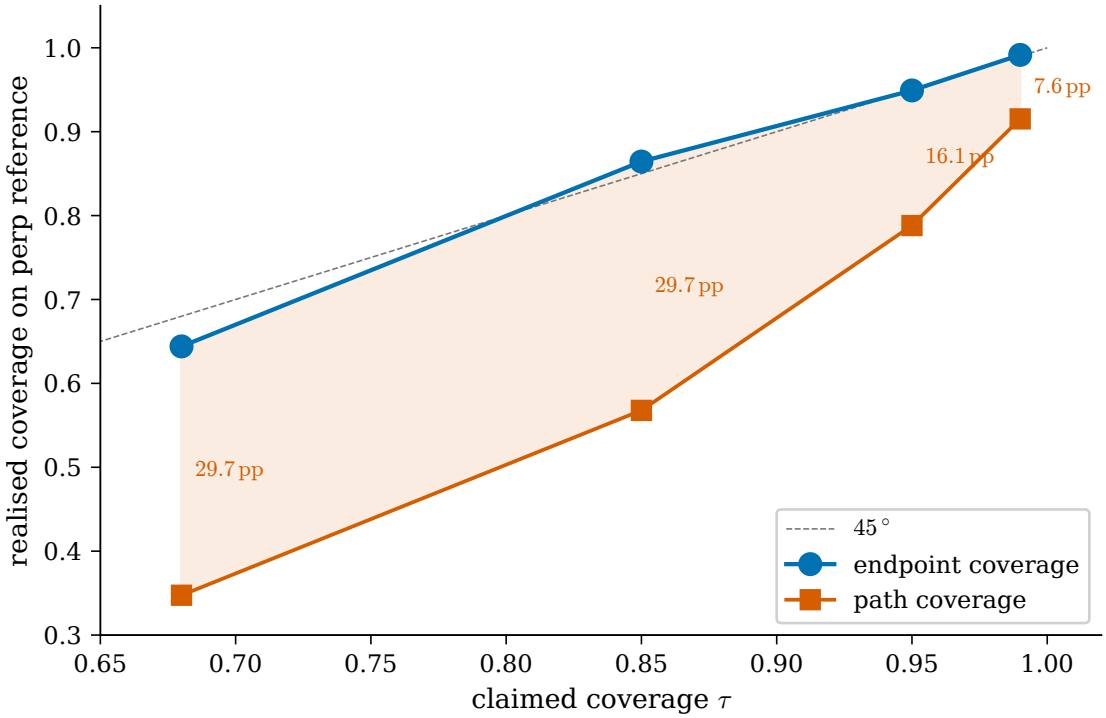

Figure 11: Path coverage gap on the 24/7 stock-perp reference. Endpoint coverage (blue) tracks nominal across $\tau \in \{0.68, 0.85, 0.95, 0.99\}$ on the perp-listed subset; path coverage (vermilion) — the fraction of weekends where the perp 1m bar high/low stay inside the served band over the full [Fri 16:00, Mon 09:30] window — runs 29.7/29.7/16.1/7.6 pp behind. The gap collapses meaningfully only at $\tau = 0.99$. Sample is small ($n = 118$ symbol-weekends across 19 calendar weekends, 2025-12 to 2026-04); the read is directional and the path-fitted conformity score (Appendix F) is the methodology fix.

Kupiec and Christoffersen pass at every served anchor — the same headline property as the weekend panel (B.4), reproduced on a different and ≈3.8× larger panel, with the OOS-fit $c(\tau)$ collapsing to ≈1.0 (the overnight bands need essentially no post-hoc widening to calibrate).

Robustness to consecutive-night autocorrelation. Overnight gaps, unlike weekends, are temporally adjacent, straining the i.i.d. assumption behind Christoffersen. A moving-block-bootstrap on the OOS violation series (block lengths $L \in \{1, 5, 10\}$, 2,000 reps) places the nominal violation rate $1 - \tau$ inside the 95% CI at every τ and every block length, and the CIs barely widen from $L = 1$ to $L = 10$ (e.g. $\tau = 0.95$: $[0.044, 0.054]$ at $L = 1$, $[0.044, 0.055]$ at $L = 10$) — the consecutive-night dependence does not invalidate the coverage claim.

Per-regime. `normal` and `high_vol` are near-nominal at every anchor (`normal` 0.691 / 0.853 / 0.951 / 0.992, $n = 4,985$; `high_vol` 0.636 / 0.836 / 0.949 / 0.992, $n = 1,405$). `earnings_night` (OOS $n \approx 60$ nights) realises 0.767 / 0.967 / 0.983 / 1.000 — over-coverage on the small-n earnings cell, the contract-favourable side of the §8 asymmetry, which tightens as the panel accumulates (§6.5).

Ex-dividend adjustment. Because the index factor does not drop for a single name's distribution, the ex-dividend-morning open is reconstructed to its cum-dividend level (`mon_open += dividend`) from scryer `yahoo/corp_actions/v1` on the 241 affected mornings. This removes the deterministic ex-div down-gap — the mean signed standardised residual on those mornings moves from -0.39 (a clear downward skew against a dividend-blind factor point) to $+0.12$, in line with the $+0.08$ panel mean — with pooled coverage unchanged. Earnings timing is session-assigned: an after-close (`amc`) release dated t_0 or a before-open (`bmo`) release dated t_1 fires inside the close(t_0)→open(t_1) gap, using the `session` field of scryer `yahoo/earnings/v2`; session timing is complete for reported earnings from 2015 onward, covering the held-out window (a date-only flag that brackets both adjacent nights dilutes the regime with normal nights and is strictly dominated).

B.16 Within-bin exchangeability — permutation test

Mondrian-CP's finite-sample coverage guarantee depends on within-bin exchangeability of the conformity score. A permutation test on the lag-1 Pearson autocorrelation of the standardised score within each (regime × symbol) bin ($n = 19$–116 per bin, 30 bins, OOS 2023+; statistic computed in `fri_ts` order, null distribution from 5,000 within-bin shuffles per bin) finds $1/30$ bins nominally rejecting at $\alpha = 0.05$ and $0/30$ at $\alpha = 0.01$ — *under-rejecting* relative to the 5% / 1% expected under exchangeability. Aggregate KS test of per-bin p-values vs Uniform(0,1) is $D = 0.276$, $p = 0.017$, but the deviation is in the under-rejecting direction (too few low p-values, too many bins with $p > 0.5$) — consistent with exchangeability holding, not violating. The single nominally-rejecting bin (GLD/normal, lag-1 $\hat{\rho} = -0.166$) does not survive Benjamini-Hochberg correction across the 30-test grid. The cross-sectional dependence of B.10 / §8 is *across* bins, not within. Source: `reports/tables/paper1_b2_exchangeability.csv`.

B.17 τ-stratified CUSUM drift bank

Beyond passive forward-tape observation (A.9), the deployed system carries a two-sided Page CUSUM drift monitor at each served τ. With control statistic $X_t = k_w/10$ (weekend violation rate), reference $\mu_0 = 1 - \tau$, and slack $k = \mu_0/2$ tuned to detect a 2× violation-rate shift, calibrated thresholds $h \in \{0.40, 0.40, 0.30\}$ for $\tau \in \{0.85, 0.95, 0.99\}$ produce mean in-control run length ≈ 225 weekends (one expected false alarm per ~ 4.3 years; calibration via 20,000 Monte-Carlo traces × 500 weekends). Detection power against an operationally relevant 2× violation-rate drift: $\sim 83\%$ with median latency 5 / 11 / 28 weekends at $\tau \in \{0.85, 0.95, 0.99\}$; against 3× drift: 2 / 5 / 14 weekends (5,000 traces × 200 weekends; `reports/tables/paper1_c2_cusum_drift.csv`, `paper1_c2_cusum_calibration.csv`).

On the 173-week OOS slice, the $\tau = 0.85$ CUSUM fires S^+ alarms at both **2023-03-10** (SVB collapse, $k_w = 7$) and **2024-08-02** (BoJ unwind, $k_w = 10$) — the two canonical stress weekends in the OOS window — with the post-BoJ S^- (over-coverage) alarm at **2024-09-27** catching the over-conservative recovery period documented under the split-anchor robustness battery (Appendix D). CUSUM banks at $\tau \in \{0.95, 0.99\}$ fire on different weekends: pairwise alarm coincidence with $\tau = 0.85$ is **2 of 7 alarms**; pairwise coincidence between $\tau = 0.95$ and $\tau = 0.99$ is **0 of 7 alarms**. Each anchor's monitor surfaces a structurally distinct aspect of drift — body-of-distribution shift at $\tau = 0.85$, near-tail at $\tau = 0.95$, deep-tail mass at $\tau = 0.99$ — confirming the τ-stratified CUSUM bank is a multi-resolution detector rather than a redundant signal. A complementary location-shift CUSUM on the standardised residual mean composes with this bank (Appendix C). Source: `reports/tables/paper1_f1_cusum_alarm_timing.csv`, `paper1_f1_cusum_alarm_proximity.csv`, `paper1_f1_cusum_alarm_coincidence.csv`. The CUSUM bank is operational concurrently with the forward-tape harness and re-uses its weekly ingest path.

C Simulation study

This appendix carries the synthetic-DGP corroboration of the per-symbol fix cited from §4.4, §6.2, and §7. Under known ground truth, it isolates the failure mechanism the $\hat{\sigma}$ standardisation closes and the one it does not.

C.1 Headline and epistemic status

Across four DGPs spanning stationary, regime-switching, drift, and structural-break specifications, the $\hat{\sigma}$-standardised architecture's per-symbol Kupiec pass-rate at $\tau = 0.95$ is 98.6–99.9% while the unweighted-Mondrian comparator's collapses to 31–38% — the per-symbol bimodality is reproducible in synthetic data with known DGP, and $\hat{\sigma}$ standardisation closes it on every DGP tested. This answers the methodological question raised by §6.2 and §7: the bimodality is a structural property of the architecture (a single per-(regime, τ) multiplier on heterogeneous-tail symbols mis-calibrates in opposing directions), not a property of the unweighted comparator on this particular real-data panel. The simulation is corroborative — a known-DGP control isolating the failure mechanism — not a pre-real-data prediction (the bimodality was first observed on real data under an unweighted Mondrian fit before the simulation was run; $\hat{\sigma}$ standardisation was implemented next; this section was committed before the real-data confirmation).

C.2 Setup

Source: `scripts/run_simulation_study.py` (~30 s end-to-end). Each DGP uses 10 synthetic symbols × 600 weekend returns, $\sigma_i \in \mathrm{linspace}(0.005, 0.030, 10)$ — spanning the empirical real-panel range. Train/OOS split at $t = 400$. Returns drawn from Student-t df=4 rescaled to $\mathrm{std}(r) = \sigma_i$ (or $\sigma_i \cdot m_t$ under DGP B/C/D's vol modifications). 100 Monte Carlo replications per DGP, seed $= 0$.

C.3 Per-DGP results

DGP	Unweighted-Mondrian pass-rate at $\tau = 0.95$	$\hat{\sigma}$-standardised pass-rate	Mechanism
A — homoskedastic baseline	0.311	**0.993**	Per-symbol scale heterogeneity is the entire failure mode
B — regime-switching vol multiplier	0.310	**0.986**	$\hat{\sigma}$ tracks regime transitions with a half-life lag
C — non-stationary scale (drift)	0.309	**0.996**	Adaptive $\hat{\sigma}$ is built for this DGP
D — variance shock (10× / 50-week transient at $t = 400$)	0.380	**0.999**	$\hat{\sigma}$ tracks even 10× variance shocks within EWMA HL=8
E — mean jump ($+\Delta = 200$ bps at $t = 400$, σ_{true} unchanged)	0.335	0.595	Location shift; $\hat{\sigma}$ absorbs Δ as variance, over-deflates low-σ scores

Both methods pool to mean realised ≈ 0.95 across DGPs A–D (within 0.005 of nominal); the split is on the *per-symbol* failure rate. The $\hat{\sigma}$-standardised architecture's smallest pass-rate (DGP B, 0.986) reflects $\hat{\sigma}$'s ~13-weekend tracking lag at regime flips. **DGP D upgrade:** the original $3\times$ variance break became a $10 \times /50$-week transient (matching real-world weekend events: 2024-08-05 BoJ unwind, March-2020 COVID-month vol multiplier $\sim 5\times$ for 6–8 weeks). LWC's per-symbol pass rate is unchanged at 99.9% (`reports/tables/paper1_c3_stronger_dgp_d.csv`).

DGPs A–D hard-code the cross-symbol scale heterogeneity that $\hat{\sigma}$ standardisation is built to remove, so they isolate the *scale* effect under known ground truth; they do not stress cross-symbol dependence or a location shift, and cannot corroborate calibration under those. DGP-E (§C.4) supplies the location-shift stress, and it is the one specification where the method's per-symbol pass-rate drops (to 59.5% at $\Delta = 200$ bps).

C.4 DGP E — location-shift limitation with closed-form phase transition

A complementary stress test — a discrete $+\Delta$ conditional-mean shift at the OOS boundary, variance unchanged — exposes a location-shift limitation that $\hat{\sigma}$-standardisation does not absorb. $\hat{\sigma}$ EWMA on the post-shift residual stream sees apparent variance $\hat{\sigma}^2 \approx \sigma^2_{\text{true}} + \Delta^2$, inflating bands across the panel and over-deflating the standardised score on symbols whose σ_{true} falls below the phase-transition threshold

$$\sigma^*(\Delta) \;\approx\; \frac{\Delta}{q \cdot c \;-\; 1}.$$

At the simulation's fitted $q_r(0.95) \cdot c(0.95) \approx 2.23$ (vs the deployed real-panel product $1.9682 \times 1.079 \approx 2.12$; §A.2), $\sigma^*(\Delta) \approx 0.81 \cdot \Delta$ — empirically validated across $\Delta \in \{100, 200, 400\}$ bps within 4% of the closed-form prediction ($\sigma^*_{\text{pred}}/\sigma^*_{\text{emp}}$ ratio in $[0.96, 0.99]$ at $\Delta \in \{100, 200\}$ bps; bounded by the σ_i grid endpoints elsewhere — `reports/tables/paper1_f2_5_sigma_star_formula.csv`). Symbols with $\sigma_{\text{true}} < \sigma^*(\Delta)$ over-cover (a sharpness deficit, the contract-favourable side per §8); symbols with $\sigma_{\text{true}} \geq \sigma^*(\Delta)$ remain calibrated. At $\Delta = 200$ bps, LWC pooled coverage over-shoots to 0.9727 and per-symbol Kupiec drops to 59.5% — the failure is *over-coverage on low-σ symbols*, not under-coverage. A complementary location-shift detector — CUSUM on the standardised residual mean, composable with the violation-rate CUSUM bank of B.17 — closes this gap at the monitoring layer, since the violation-rate CUSUM is by-construction blind to location shifts that produce nominal violation rates. This is the "location shifts" bound listed in §8. Source: `reports/tables/paper1_f2_mean_jump_bimodality.csv`.

C.5 Sample-size sweep and the symbol-admission floor

A sample-size sweep (`scripts/run_simulation_size_sweep.py`; 7 N-values × 4 DGPs × 100 reps × 2 forecasters = 22,400 cells) establishes the newly-listed-symbol admission threshold: $N \geq 80$ weekends under stationary / drift / structural-break DGPs, $N \geq 200$ under regime-switching (the production threshold). HOOD's empirical

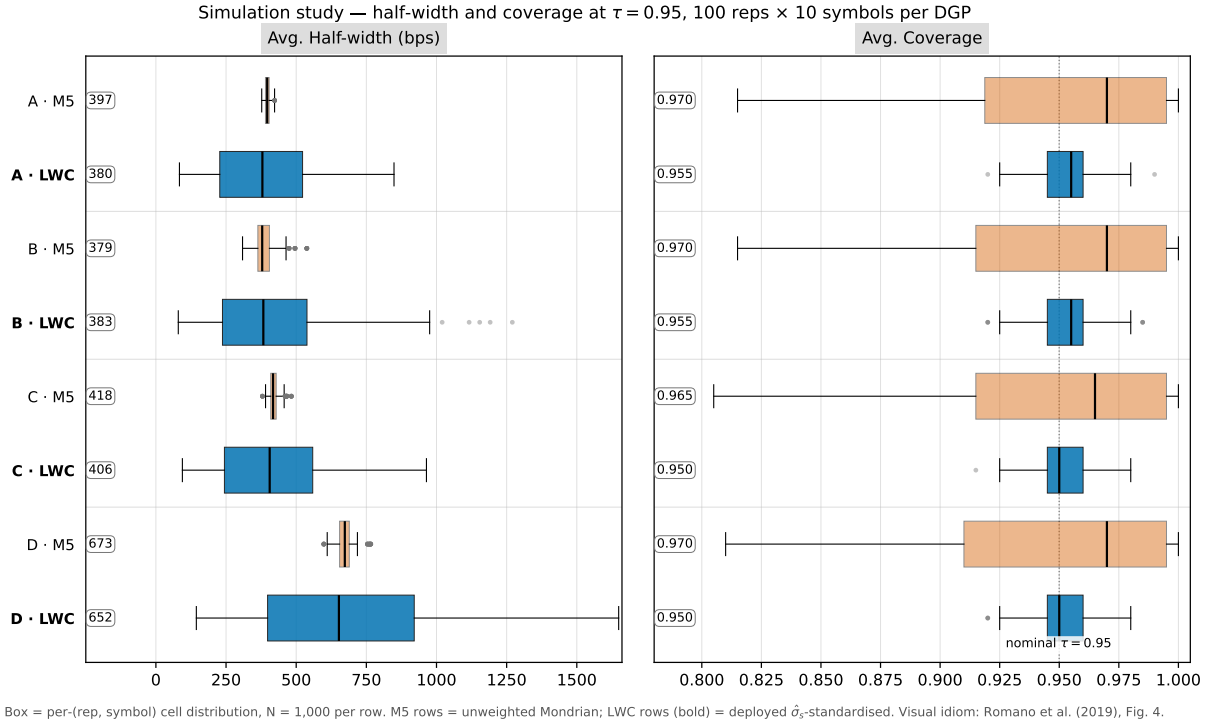

Figure 12: Simulation study at $\tau = 0.95$ — half-width and coverage in the visual idiom of [Romano et al., 2019] (their Fig. 4), adapted to the per-symbol Mondrian setting. Eight rows: 4 DGPs $\times$ 2 forecasters (M5 unweighted-Mondrian comparator in vermilion; LWC deployed $\hat{\sigma}$-standardised in blue, bold). Each box is the per-(rep, symbol) cell distribution ($N = 1,000$ per row). Median pills annotate the leftmost edge of each box. Left panel: avg. half-width (bps). Right panel: avg. coverage; dotted vertical marks nominal $\tau = 0.95$. The $\hat{\sigma}$ contribution is the *coverage panel*: M5 boxes scatter from ~ 0.81 to ~ 1.00 — the per-symbol bimodality documented in §7 and Appendix D — while LWC boxes sit tightly on $\tau = 0.95$ across every DGP. The half-width panel makes the Appendix D redistribution visible: M5 widths concentrate on a single per-rep value (the unweighted multiplier applied uniformly across the σ-grid); LWC widths span a wide range within each rep (per-symbol $\hat{\sigma}_s$ scaling).

$N \approx 238$ (246 listed $-$ 8 $\hat{\sigma}$ warm-up) sits inside this band; HOOD's empirical Kupiec $p = 0.329$ at $\tau = 0.95$ (violation rate 0.035, slight over-coverage; B.11) is consistent with the simulation's pass-rate range at $N = 200$. See `reports/m6_simulation_study.md` and `reports/tables/sim_size_sweep_admission_thresholds.csv`.

Operationally, the $\hat{\sigma}$ warm-up rule ($\geq$8 past observations, §4.3) sets only the empirical floor; the sweep extends it to a panel-admission threshold. A protocol admitting a symbol with fewer than 200 weekends of history should treat the per-symbol calibration claim as deferred until the panel accumulates.

C.6 Summary figure

D Ablation detail and the $\hat{\sigma}$ ladder

§7 carries the effect sizes for the three load-bearing components; this appendix carries the full tables and equations, the walk-forward $\delta(\tau)$ finding, the c(τ) schedule analysis, the $\hat{\sigma}$ selection ladder and its supporting robustness battery, the vol-index switchboard selection evidence, the width-redistribution decomposition, and the tokenized-tracking baseline — whose deployment headline (-45% half-width, -46% Winkler at matched coverage; 7 of 9 symbols) appears in §6.2, with the construction, full grid, and the matched-calibration-history decomposition (D.7.7, which shows about half that edge is the 12-year calibration record rather than the architecture) here.

All deltas carry block-bootstrap 95% CIs by weekend (1,000 resamples, seed=0, paired by `(symbol, fri_ts)`). Raw tables: `reports/tables/v1b_constant_buffer_*.csv`, `reports/tables/v1b_mondrian_*.csv`, `reports/tables/m5_vs_m6_bootstrap.csv`, `reports/tables/sigma_ewma_*.csv`, `reports/tables/paper1_c2_tokenised_tracking_baseline*.csv`.

D.1 Regime stratification — vs constant-buffer baseline

The constant-buffer baseline is the Pyth/Chainlink Friday close held forward, with a single global symmetric buffer $b(\tau)$ per claimed quantile — one parameter per τ, no factor switchboard, no residual quantile, no regime model:

$$\left[\, p_{\text{Fri}}\,(1 - b(\tau)),\ p_{\text{Fri}}\,(1 + b(\tau))\,\right], \qquad b(\tau) \text{ calibrated globally on the training panel.}$$

D.1.1 Trained-buffer fit on OOS

Each $b(\tau)$ is the empirical τ-quantile of $|p_{\text{Mon}} - p_{\text{Fri}}|/p_{\text{Fri}}$ on the calibration set. Carrying the trained $b(\tau)$ to the 2023+ panel:

τ	method	realised	half-width (bps)	p_{uc}	p_{ind}
0.68	constant buffer (train-fit)	0.538	76.0	0.000	0.001
0.68	deployed	0.693	130.8	0.264	0.244
0.85	constant buffer (train-fit)	0.731	140.0	0.000	0.000
0.85	deployed	0.855	213.6	0.565	0.403
0.95	constant buffer (train-fit)	0.897	272.0	0.000	0.013
0.95	deployed	0.950	370.6	0.956	0.603
0.99	constant buffer (train-fit)	0.984	695.0	0.018	0.927
0.99	deployed	0.990	635.0	0.942	≈ 1.0

The training-fit constant buffer under-covers at every $\tau \leq 0.95$ (deficits $-14.2/-12.0/-5.4$pp, all rejecting Kupiec at $p_{uc} < 10^{-6}$). The mechanism is non-stationarity: $b(\tau)$ is calibrated on a 2014–2022 window calmer than the 2023+ holdout, and a single global scalar has no channel through which to adapt. The deployable constant-buffer baseline does not deliver coverage on the holdout; the regime-stratified architecture does.

D.1.2 Coverage-matched comparison

To answer the width-at-coverage question, we re-fit $b(\tau)$ post-hoc on the OOS slice (oracle-fit, non-deployable, but gives the constant buffer the most generous trade-off):

τ	matched b	matched-CB hw (bps)	deployed hw (bps)	Δ width (deployed − CB)
0.68	1.21%	120.9	130.8	**+8.2% [+3.7, +13.0]**
0.85	2.26%	226.4	213.6	**−5.7% [−10.6, −0.5]**
0.95	3.96%	395.6	370.6	**−6.3% [−12.1, −0.4]**
0.99	5.46%	545.8	635.0	+16.3% [+9.6, +23.6]

The deployed architecture is narrower than a coverage-matched (oracle-fit) constant buffer at $\tau \in \{0.85, 0.95\}$. The $\hat{\sigma}$ standardisation that redistributes width across symbols (§D.6) reaches sufficient sharpness on the equity-heavy panel that the deployment outperforms even the oracle-fit constant buffer at the headline anchor; the architecture is wider only at $\tau \in \{0.68, 0.99\}$, where the cell-conditional regime structure carries genuine information. The regime model + $\hat{\sigma}$ buy *both* coverage on a non-stationary distribution *and* mean-width sharpness at $\tau \in \{0.85, 0.95\}$ against the strongest non-modelled comparator.

D.1.3 Per-regime decomposition

The pooled half-width at $\tau = 0.95$ hides a regime-conditional structure. Deployed pooled: 370.6 bps; per-regime $300.1/334.7/603.7$ bps (normal / long_weekend / high_vol). The matched-CB at $\tau = 0.95$ delivers 395.6 bps in *every* regime by construction (one global multiplier). The regime model concentrates width in `high_vol` (+52% vs CB at

603.7 vs 395.6) where it earns -1.1pp coverage; releases width in `normal` (-24% vs CB at 300.1 vs 395.6) without losing coverage. The constant buffer's pooled coverage is average-right but worst where it matters most: violations concentrate in high-vol weekends — exactly where a downstream lending protocol incurs the largest mark-to-market gap. The regime model's contribution is **(i) calibration through distribution shift** and **(ii) tail-protection in `high_vol` weekends** at sharper *normal*-regime widths than the constant buffer.

D.2 Per-symbol $\hat{\sigma}$ standardisation — vs unweighted Mondrian

The *unweighted Mondrian comparator* shares the regime structure of D.1 but omits the per-symbol scale standardisation: a single per-(regime, τ) conformal quantile is fit on the *un-standardised* relative residual $|P_{\text{Mon}} - \hat{P}_{\text{Mon}}|/P_{\text{Fri}}$, and the served band uses that quantile directly as a fraction of P_{Fri}. Same training set, same OOS slice, same point estimator, same regime classifier, same finite-sample CP rank formula. Only the $\hat{\sigma}$ standardisation of the conformity score is added by the deployed architecture.

D.2.1 OOS pooled results

τ	method	realised	hw (bps)	p_{uc}	p_{ind}	per-symbol Kupiec pass
0.68	unweighted Mondrian	0.735	137.5	0.000	0.339	(1/10)
0.68	deployed	0.693	130.8	0.264	0.244	10/10
0.85	unweighted Mondrian	0.887	235.1	0.000	0.424	(1/10)
0.85	deployed	0.855	213.6	0.565	0.403	10/10
0.95	**unweighted Mondrian**	**0.950**	**354.6**	**0.956**	**0.921**	**2/10**
0.95	**deployed**	**0.950**	**370.6**	**0.956**	**0.603**	**10/10**
0.99	unweighted Mondrian	0.990	677.7	0.942	0.344	(8/10)
0.99	deployed	0.990	635.0	0.942	≈ 1.0	10/10

Both methods pool to nominal at $\tau = 0.95$ (both 0.950, both Kupiec $p = 0.956$). The split is on per-symbol calibration: the unweighted Mondrian comparator passes 2 of 10 per-symbol Kupiec cells at $\tau = 0.95$; the deployed architecture passes 10 of 10. The unweighted comparator pools to nominal *through compensating per-symbol biases* — heavy-tail tickers under-cover, low-vol tickers over-cover, and the cross-sectional average lands at τ by accident. The $\hat{\sigma}$ standardisation factors out the per-symbol scale before the conformal fit, so the regime quantile carries only the *shape* of the standardised residual distribution and per-symbol calibration is recovered by the per-symbol $\hat{\sigma}_s(t)$ at serve time.

The pooled half-width tax at $\tau = 0.95$ is $+4.5\%$ (370.6 vs 354.6 bps; $+15.93$ bps with 95% block-bootstrap CI $[+1.05, +30.73]$ from `m5_vs_m6_bootstrap.csv`). The per-symbol calibration is bought at exactly that price; §D.6 documents how the +4.5% pooled figure resolves into a +17.8% widening on heavy-tail equities and a -48.8% narrowing on the defensive class.

D.2.2 Side-by-side at $\tau = 0.95$

The $\hat{\sigma}$-standardisation isolation: same point estimator, same regime classifier, same Mondrian split-conformal, same finite-sample CP rank formula, same training set, same OOS slice. Only the $\hat{\sigma}$ standardisation of the conformity score is added.

metric	unweighted Mondrian	deployed ($\hat{\sigma}$-standardised)
pooled realised	0.9503	0.9503
pooled half-width (bps)	354.6	370.6
per-symbol Kupiec pass-rate	2 / 10	**10 / 10**
per-symbol Berkowitz LR range	0.9 – 224 (250×)	**3.2 – 16.7 (5.2×)**
LOSO realised std at $\tau = 0.95$	0.0759	**0.0128 (5.9× tighter)**

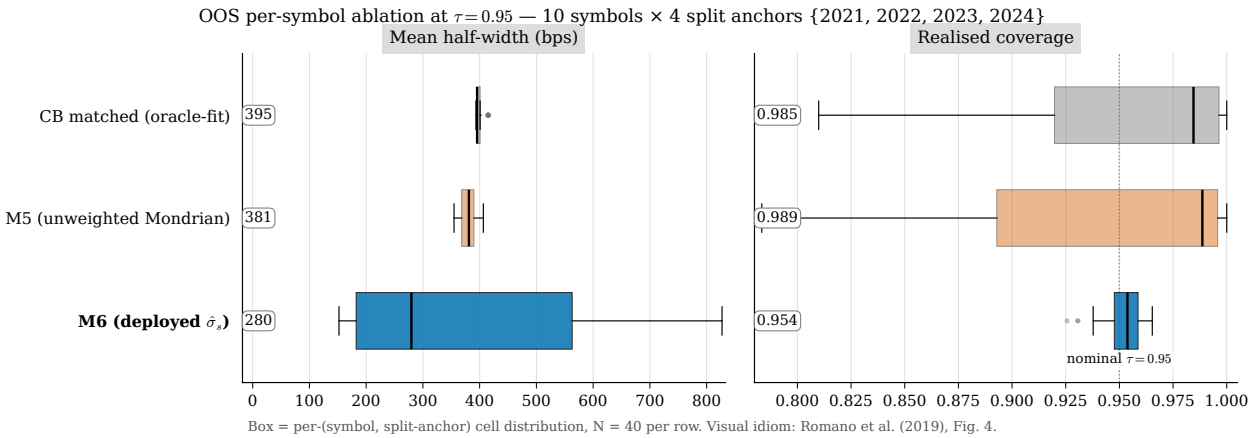

Figure 13: Per-symbol ablation at $\tau = 0.95$ on real data, in the visual idiom of [Romano et al., 2019] (their Fig. 4) adapted to the per-symbol Mondrian setting. Three method rows — constant-buffer (oracle-fit on the OOS slice; the most generous comparator D.1.2 considers), unweighted Mondrian (M5; same architecture as deployed without $\hat{\sigma}$ standardisation), and the deployed $\hat{\sigma}$-standardised architecture (M6; bold). Each box is the per-(symbol, split-anchor) cell distribution: 10 symbols $\times$ 4 split anchors $\{2021, 2022, 2023, 2024\}$-01-01 = 40 cells per row. Median pills annotate the leftmost edge. Left panel: mean half-width (bps). Right panel: realised coverage; dotted vertical marks $\tau = 0.95$. The deployed-$\hat{\sigma}$ row collapses per-symbol coverage onto nominal (cross-cell std 0.0092) while CB-matched and M5 disperse from ~ 0.78 to 1.00 (0.071 and 0.062 respectively) — pooled-coverage parity through compensating per-symbol biases. The half-width panel is the dual: M6 widths span $\sim$ 150–830 bps within each split anchor (per-symbol $\hat{\sigma}_s$ redistribution; §D.6), while CB and M5 are uniform across symbols by construction. Source table: `reports/tables/paper1_fig7b_per_symbol_ablation.csv`.

metric	unweighted Mondrian	deployed ($\hat{\sigma}$-standardised)
LOSO Kupiec pass-rate (held-out)	8 / 10	**10 / 10**

The per-symbol Kupiec pass-rate, Berkowitz LR range, and LOSO calibration std all improve simultaneously by mechanism — no other component of the architecture is touched. The Appendix C simulation evidence on synthetic data with known DGP (data-generating process) confirms the same trade under ground truth — under four DGPs spanning homoskedastic, regime-switching, drift, and structural-break specifications, the unweighted Mondrian comparator's per-symbol Kupiec pass-rate at $\tau = 0.95$ collapses to 31–38% while $\hat{\sigma}$ standardisation closes the bimodality on every DGP at 98.6–99.9%.

D.2.3 Walk-forward $\delta(\tau)$ selection — collapses to zero

Under an unweighted Mondrian fit, a plain $c(\tau)$ correction to the pooled OOS slice produces per-split realised coverage that scatters around nominal — it passes Kupiec on the pooled fit by construction but under-covers in roughly half the splits, and stabilising it requires a per-anchor structural-conservatism shift $\delta(\tau)$. Under the $\hat{\sigma}$-standardised architecture the same walk-forward sweep over $\delta \in \{0.00, \ldots, 0.07\}$ on the four powered tune fractions $f \in \{0.40, 0.50, 0.60, 0.70\}$ selects $\delta(\tau) \equiv 0$: per-symbol scale standardisation tightens cross-split realised-coverage variance enough that the shift is not load-bearing. The δ schedule an unweighted comparator requires thus collapses to identity once $\hat{\sigma}$ is per-symbol. The schedule is retained as a four-zero vector in the artefact JSON for shape-compatibility with the receipt schema, and the deployment carries zero walk-forward-tuned scalars (§4.5).

D.3 Near-identity OOS $c(\tau)$ bump

The deployed $c(\tau)$ schedule is

$$c(\tau) = \{0.68{:}1.000,\ 0.85{:}1.000,\ 0.95{:}1.079,\ 0.99{:}1.003\}.$$

This is a 4-scalar multiplicative correction fit on the 2023+ OOS slice as the smallest $c \in [1, 5]$ such that pooled realised coverage with effective quantile $c \cdot q_r(\tau)$ matches τ. Three of the four are essentially identity: under the deployed $\hat{\sigma}$ rule and the 2023-01-01 split, the trained per-regime quantile already lands within 0.3 pp of nominal at

$\tau \in \{0.68, 0.85, 0.99\}$, so $c(\tau)$ at those anchors collapses to $\{1.000, 1.000, 1.003\}$. Only $\tau = 0.95$ has a meaningful train-OOS distribution-shift gap; the 7.9% widening at that anchor is what closes the pooled $0.946 \to 0.950$ correction.

$c(\tau)$ thus carries one scalar of meaningful OOS information, not four. Dropping the $c(\tau)$ correction and serving at the trained-quantile coverage costs ~0.4 pp at $\tau = 0.95$ on this slice (0.946 vs the deployment-tuned 0.950) — statistically indistinguishable from nominal under Kupiec, but a persistent deficit on the served claim. §8 carries the provenance disclosure for $c(0.95)$; §6.1's LOSO cross-validation re-validates it on held-out symbol data.

D.4 $\hat{\sigma}$ selection procedure and multiple-testing disclosure

The per-symbol scale rule $\hat{\sigma}_s(t)$ is the only deployment constant of the architecture without a closed-form derivation. The comparison set is a five-variant ladder on the 2023+ OOS slice — K=26 trailing window (baseline), EWMA HL=$\{6, 8, 12\}$, and a 50/50 K=26 / EWMA-HL-8 convex blend — under the pre-registered three-gate criterion stated in §7: **Gate 1**, no per-cell rejection at uncorrected $\alpha = 0.05$ on the 16-cell split-date Christoffersen grid; **Gate 2**, per-symbol Kupiec 10/10 at $\tau = 0.95$; **Gate 3**, 95% bootstrap CI upper bound on Δhw% (vs K=26) within $+5\%$ at every τ. All five variants share the $\geq$8-past-observation warm-up and pre-Friday convention; per-variant artefacts (q_r, $c(\tau)$, $\delta(\tau)$) were re-fit at the same 2023-01-01 cutoff with δ collapsing to zero on every variant.

Gate 1 (multi-test corrected) does not discriminate. At uncorrected α, EWMA HL=8 is the only variant with zero per-cell rejections; K=26 carries 4. Under Benjamini–Hochberg (false-discovery-rate) correction at FDR=0.05 across the full 80-cell grid (5 variants × 16 cells), no variant has any rejected cell — the smallest q-value in the grid is 0.130 (K=26, 2022 split, $\tau = 0.95$). After multi-test correction, per-cell Christoffersen evidence does not statistically distinguish the five variants. **Gate 2 (per-symbol Kupiec) also does not discriminate** — all five pass 10/10 at $\tau = 0.95$, which the D.2 ablation already established as a property of per-symbol $\hat{\sigma}$ standardisation rather than of the rule's half-life.

Gate 3 (bootstrap CI on width) is load-bearing (`reports/tables/sigma_ewma_bootstrap.csv`):

τ	EWMA HL=8 Δhw% (vs K=26)	95% CI on Δhw%	within +5% gate?
0.68	-1.34%	$[-3.02, +0.25]$	✓
0.85	-2.07%	$[-3.75, -0.45]$	✓ (also significantly narrower)
0.95	-3.83%	$[-6.15, -1.88]$	✓ (also significantly narrower)
0.99	-7.41%	$[-9.33, -5.65]$	✓ (also significantly narrower)

EWMA HL=8 is narrower than K=26 at every τ in point estimate and statistically significantly narrower at $\tau \in \{0.85, 0.95, 0.99\}$; calibration is preserved (paired Δ-realised CIs straddle zero, per-symbol Kupiec 10/10 holds). Gate 3 has no multi-test exposure: one bootstrap CI per τ, threshold pre-registered. EWMA HL=8 satisfies all three gates and delivers the largest narrowing relative to baseline.

Held-out re-validation. The forward-tape harness (§5.8) carries a sibling variant-comparison runner (`scripts/run_forward_tape_variant_comparison.py`) that loads a content-addressed *variant bundle* with frozen $(q_r, c(\tau), \delta(\tau))$ schedules for all five variants and applies them to forward weekends as they accumulate. The runner *never re-selects* — its function is to flag if a different $\hat{\sigma}$ rule looks dramatically cleaner on truly held-out data. Status at submission: $N = 11$ forward weekends (`reports/m6_forward_tape_11weekends_variants.md`, 2026-05-01 $\to$ 2026-07-10); N has cleared the harness's own preliminary banner ($N \geq 4$) though moderate per-variant power still requires $N \geq 13$. On the forward slice the five variants are near-indistinguishable — pooled half-widths at $\tau = 0.95$ span 293.6–321.5 bps with identical realised coverage (0.964), and no variant looks dramatically cleaner than canonical EWMA HL=8, so the re-validation raises no flag to revisit the selection. The $\hat{\sigma}$ deployment claim therefore rests on Gate 3 plus the (accumulating) forward-tape re-validation. We do not claim the $\hat{\sigma}$ rule is optimal among all locally-weighted variants (§3.5 non-goal).

D.4.1 Regime quartile cutoff q — ablation under the three-gate frame

The deployed `high_vol` regime gate (§5.5) is "VIX at Friday close in the top quartile of its trailing 252-trading-day window" — a single scalar $q = 0.75$. Ablating $q \in \{0.60, 0.67, 0.70, 0.75, 0.80, 0.90\}$ under the three-gate frame of D.4 (`reports/tables/paper1_b1_regime_threshold_ablation.csv`, `..._hw_bootstrap.csv`):

q	high_vol mix (%)	Gate 1: pooled Kupiec all τ	Gate 1b: pooled Christoffersen all τ	Gate 2: per-symbol Kupiec	Δhw% at $\tau = 0.95$ vs deployed (95% CI)
0.60	38.9	✓	✓	10/10	-2.76% $[-3.80, -1.73]$
0.67	30.8	✓	✓	10/10	-1.78% $[-2.66, -0.87]$
0.70	28.8	✓	✓	10/10	$+1.73\%$ $[+0.85, +2.71]$
0.75 (deployed)	**23.9**	✓	✓	**10/10**	—
0.80	21.1	✓	✓	10/10	-1.70% $[-2.11, -1.32]$
0.90	12.3	✓	✗ rejects	10/10	-4.30% $[-5.34, -3.25]$

5 of 6 candidates satisfy Gates 1, 1b, and 2. Three alternates ($q \in \{0.60, 0.67, 0.80\}$) deliver 1.7–2.8% narrower bands at preserved calibration with bootstrap CIs excluding zero. The deployed $q = 0.75$ is **convention-anchored** (top quartile by definition) rather than width-optimization-selected. The differences are operationally small (~ 5 bps on a 370 bps $\tau = 0.95$ headline) but real; a re-tuning on a held-out tune slice could close this gap.

D.4.2 Split anchor — robustness across {2021, 2022, 2023, 2024}

The deployed train/test split anchor is 2023-01-01 (1 scalar). Ablating across $\{2021, 2022, 2023, 2024\}$-01-01 (re-fit per anchor; `reports/tables/m6_lwc_robustness_split_sensitivity.csv`):

split anchor	n_{OOS} weekends	$\tau = 0.95$ realised	Kupiec p	Christoffersen p	hw bps	per-sym Kupiec
2021-01-01	277	0.9539	0.348	0.115	357.2	10/10
2022-01-01	225	0.9520	0.661	0.186	363.5	10/10
2023-01-01 (deployed)	**173**	**0.9503**	**0.956**	**0.603**	**370.6**	**10/10**
2024-01-01	121	0.9504	0.947	0.671	416.8	10/10

Realised $\tau = 0.95$ coverage stays in $[0.9503, 0.9539]$; Kupiec $p \in [0.35, 0.96]$; Christoffersen $p \in [0.12, 0.67]$; per-symbol Kupiec 10/10 across all four anchors. The deployed split anchor is robust. Half-width varies (357 → 417 bps) — driven by an eval-slice composition effect: later eval slices contain more high-$\hat{\sigma}$ weekends, including the 2024-08-05 BoJ unwind (§8).

D.4.3 Cross-asset regime-index sensitivity

The classifier uses VIX as the high-vol gate for *all* symbols, including GLD (gold) and TLT (long-dated treasury) which have asset-specific vol indices (GVZ, MOVE) used in the $\hat{\sigma}$ regression (§5.4). A sensitivity ablation that swaps GLD's regime gate to GVZ and TLT's to MOVE — flipping the regime tag on 23% of GLD weekends and 28% of TLT weekends — leaves the pooled $\tau = 0.95$ headline coverage unchanged at 0.9503 (Kupiec $p = 0.956$), per-symbol Kupiec 10/10, and half-width within 0.1%. The $\hat{\sigma}$-standardisation absorbs the regime-tagging difference structurally; the regime cell contributes a small marginal adjustment via $q_r(\tau)$, not a load-bearing scale separation. Source: `reports/tables/paper1_b4_regime_index_sensitivity.csv`, `..._per_symbol.csv`. The deployed VIX-only classifier is justified by simplicity and per-asset-vol-index-agnosticism, not by being optimally tuned per asset.

D.5 Vol-index switchboard selection evidence

The per-symbol vol-index assignments of §5.4 (GVZ for GLD, MOVE for TLT, VIX elsewhere) are evidence-driven rather than conventional. In the V1b validation pass, fitting the log-log sigma regression with VIX as the vol index

yielded $\hat{\beta} \approx 0.55$ for GLD and $\hat{\beta} \approx 0.94$ for TLT, well below the equity-class mean ($\hat{\beta} \approx 1.5$) — the VIX level carries materially less information about defensive-class weekend residual scale than about equity scale. Substituting GVZ and MOVE lifted $\hat{\beta}$ into the equity range and improved coverage at matched bandwidth. Note the asymmetry with D.4.3: the asset-native indices are load-bearing in the *scale* pathway but not in the *regime-gate* pathway, where $\hat{\sigma}$ standardisation absorbs the tagging difference.

D.6 How $\hat{\sigma}$ standardisation redistributes width across symbols

D.1.3 attributed the constant-buffer premium to `high_vol`; D.2 attributed the per-symbol Kupiec fix to $\hat{\sigma}$ standardisation; this section attributes the +4.5% pooled half-width tax to a width *redistribution* across symbols within a regime.

Per-class decomposition at $\tau = 0.95$. Where the unweighted Mondrian comparator pays a uniform width within a regime, the deployed architecture redistributes that width across symbols within a regime via the per-symbol $\hat{\sigma}_s$ multiplier:

class	n	unweighted Mondrian hw at $\tau = 0.95$	deployed hw at $\tau = 0.95$	Δ
equities (8 symbols)	1,384	354.6 bps	417.8 bps	**+17.8%** (carries the per-symbol heavy-tail fix)
GLD + TLT (2 symbols)	346	354.6 bps	181.6 bps	**−48.8%** (releases over-conservative defensive bands)

Row-weighted reconciliation against the §6.1 panel pooled: $(1384 \cdot 417.8 + 346 \cdot 181.6)/1730 = 370.6$ bps, matching the §6.1 panel-pooled half-width at $\tau = 0.95$.

Under the unweighted comparator, every symbol within a regime received the same $c(\tau) \cdot q_r(\tau) \cdot p_{\text{Fri}}$ width — the common-multiplier-on-heterogeneous-tails failure visible in Fig. S1 (§6.2); the per-class mean HW is identical across classes because the regime mix is identical across symbols (regime is global, §5.5) and width within a regime carries no per-symbol axis. Under the deployed architecture, the $\hat{\sigma}$ multiplier is per-symbol, so heavy-tail equities (HOOD, MSTR, TSLA) receive bands that match their relative residual scale, and the over-covered defensive class (GLD, TLT) receives narrower bands without losing coverage.

The +4.5% pooled width tax of D.2.1 is the *equity-weighted* read on a panel that is 80% equity rows ($1,384/1,730$). Under the deployed architecture the average equity-class symbol receives a +17.8% wider band than under an unweighted Mondrian fit (417.8 vs 354.6 bps); a defensive-class symbol (GLD, TLT) receives a −48.8% narrower one (181.6 vs 354.6). The redistribution is *toward heavy-tail equities*, where the per-symbol calibration evidence demands it (§6.2 / Fig. S1: TSLA, MSTR, HOOD all in the 0.035–0.046 violation-rate band that the unweighted comparator could not deliver), and *away from defensive collateral*, where the regime-pooled multiplier was over-conservative. A protocol whose collateral mix is dominated by equity xStocks should expect the deployed mean served band to be materially wider than the unweighted comparator's; a protocol whose collateral mix is dominated by GLD/TLT will see materially narrower bands. The pooled +4.5% figure is the right summary of the trade only at the deployed panel composition; downstream consumers should re-weight by their own collateral mix.

The residual these components do not reach — within-weekend cross-sectional common-mode, and the per-symbol Berkowitz rejections on TSLA/TLT — is §8's subject; a characterised (not deployed) cluster-conditional variant is in Appendix F.

D.7 Tokenized-tracking baseline (post-Cong)

§6.3 carries the headline of this comparison; this section carries the construction, the full snapshot and per-symbol grids, and the power caveat.

The D.1 constant-buffer baseline is the *underlying-frozen* competitor archetype — what a consumer reading the Pyth or Chainlink `price` field gets during the closed window. The complementary archetype, motivated empirically by Cong et al. [Cong et al., 2025] (Table 10, p. 40), is the *tokenised-tracking* baseline: an oracle that publishes point_t = tokenised-side price at t throughout the closed window. Cong et al. report a Hasbrouck-style passthrough $\lambda = 0.903$ with adjusted $R^2 = 0.839$ from off-hour tokenised returns to the close-to-open underlying return — i.e.,

reading the live tokenised perp during closed hours already explains 84% of the variance in where the underlying reopens. The tokenised-tracking critique therefore has empirical bite on the conditional mean and must be addressed on the conditional quantile. It is a *constructed proxy* for the continuous-tracking archetype, not any vendor's product, and is distinct from D.1's underlying-frozen archetype and from §6.2 / D.2's internal-architecture comparators.

D.7.1 Baseline construction

For each canonical snapshot t in the closed window:

$$\text{point}_t = \text{perp\_close}_{\text{kraken\_futures}}(t), \qquad \text{halfwidth}_{t,\tau} = \text{quantile}_\tau\Big(\big|\text{mon\_open} - \text{perp\_close}(t)\big| \;;\; \text{past weekends pooled across symbols}\Big)$$

Walk-forward expanding-window calibration over weekends (warm-up = 4 past weekends before the first evaluable observation). The empirical-quantile residual specification is intentional: it places the baseline in the same non-parametric quantile family as the deployed architecture, eliminating the objection that the baseline is handicapped by an imposed Gaussian — the objection §6.2's GARCH baselines invite.

Snapshot grid. Six canonical evaluation moments in [Fri 16:00 ET, Mon 09:30 ET]: `fri_close` (Fri 16:00 ET), `sat_noon` (Sat 12:00 ET), `sun_noon` (Sun 12:00 ET), `sun_globex` (Sun 20:00 ET CME Globex reopen), `mon_premkt` (Mon 04:00 ET pre-market start), `mon_open` (Mon 09:00 ET just before NMS — national market system / regular-session — reopen). The deployed M6 LWC band is published once at Friday close and held constant; the baseline updates at every snapshot. The baseline's `mon_open` snapshot is its most-informed configuration — the tokenised side has absorbed the entire closed-window news flow by then.

D.7.2 Panel

The post-launch slice of the LWC artefact (fri_ts $\geq$ 2025-12-19) intersected with the `cex_stock_perp/ohlcv/v1` kraken_futures xstock-backed perp tape: $n = 117$ weekend-symbol observations pre-warmup, $n = 105$ after warm-up. Nine symbols (SPY, QQQ, GLD: 19 weekends each; TSLA, NVDA, GOOGL, AAPL, HOOD, MSTR: ≈ 10 each). **TLT is excluded** — no xstock-backed perp exists for it. This is materially smaller than the $n = 1,730$ panel that backs §6 and D.1–D.6, and §6.2's grid backtests are the powered baselines; D.7 is a head-to-head on a powered-only-for-paired-Winkler post-launch sub-slice. See `reports/v1b_tokenised_tracking_baseline.md` for the full panel description.

D.7.3 Head-to-head — pooled

The deployed M6 LWC band (one $(point, halfwidth)$ per weekend, held across the closed window) on the bake-off panel; tokenised-tracking baseline at its mon_open snapshot (the most-informed configuration). Cong et al.'s $\lambda = 0.903$ channels through this row. Winkler is the interval score glossed in §6.3 — width plus a penalty for misses:

τ	method	realised	hw (bps)	Winkler (bps)	Kupiec p_{uc}
0.68	tokenised-tracking (mon_open)	0.724	205	661	0.330
0.68	**this paper (M6 LWC)**	**0.641**	**124**	**426**	**0.371**
0.85	tokenised-tracking (mon_open)	0.867	372	1,031	0.627
0.85	**this paper (M6 LWC)**	**0.863**	**207**	**588**	**0.684**
0.95	tokenised-tracking (mon_open)	0.943	656	1,619	0.742
0.95	**this paper (M6 LWC)**	**0.949**	**358**	**879**	**0.949**
0.99	tokenised-tracking (mon_open)	0.981	904	2,572	0.407

τ	method	realised	hw (bps)	Winkler (bps)	Kupiec p_{uc}
0.99	**this paper (M6 LWC)**	**0.991**	**636**	**1,516**	**0.871**

At $\tau = 0.95$ the deployed architecture matches the baseline's coverage rate (0.949 vs 0.943, both Kupiec-passing) at 45% **narrower half-width** (358 vs 656 bps) and 46% **lower Winkler interval score** (879 vs 1,619 bps). At $\tau = 0.99$ the Winkler gap widens to -41% — the tail regime where the baseline's empirical-quantile fattens dramatically. Across all six snapshots and four served τ (24 baseline cells), the deployed architecture's mean Winkler is lower on 24 of 24.

D.7.4 Snapshot evolution across the closed window

The strongest configuration a tokenised-tracking competitor can claim is its latest snapshot — by Monday pre-open the tokenised perp has absorbed the weekend news flow. The snapshot evolution at $\tau = 0.95$ measures how much that helps:

snapshot	baseline hw (bps)	baseline Winkler (bps)	Winkler ratio vs M6 LWC
`fri_close`	730	1,685	1.92× worse
`sat_noon`	733	1,702	1.94× worse
`sun_noon`	735	1,698	1.93× worse
`sun_globex`	736	1,696	1.93× worse
`mon_premkt`	691	1,688	1.92× worse
`mon_open`	**656**	**1,619**	**1.84× worse**

The baseline tightens by $\sim 10\%$ (730 $\rightarrow$ 656 bps) across the closed window, but never closes more than half the gap to the deployed architecture's flat 358 bps. Cong et al.'s λ = 0.903 transports the conditional mean; it does not transport the residual distribution, which is what consumes width.

A 15-minute-resolution per-minute Winkler curve (`reports/tables/paper1_c3_per_minute_winkler_curve.csv`; runner `scripts/run_v1b_tokenised_per_minute_winkler.py`) sharpens the finding: at $\tau = 0.95$ the baseline's Winkler is **worst** during Saturday night through Sunday afternoon (hour offsets 32–44 past Fri 16:00 ET, mean Winkler $\approx 2,023$ bps; ratio $2.14\times$ the deployed flat ≈ 945 bps). Waiting until Sunday night for the perp to absorb weekend news does not help — Sunday afternoon is the *worst*-scoring time on the perp, consistent with crypto-native weekend flow drifting the perp price while no underlying-market arbitrage operates to correct it. The Winkler ratio is monotone in neither direction across the closed window; the baseline never closes more than ∼half the gap at any minute.

D.7.5 Per-symbol robustness — where the baseline is competitive

The pooled comparison is honest about where the result is and is not unanimous. Per-symbol Winkler ratio (baseline / M6 LWC at $\tau = 0.95$, `mon_open` snapshot; > 1 means M6 LWC wins):

symbol	n	M6 LWC hw (bps)	baseline hw (bps)	Winkler ratio
HOOD	10	635	2,109	**3.32**
GOOGL	10	348	509	**2.91**
NVDA	10	399	858	**2.15**
MSTR	10	735	1,153	**1.57**
GLD	15	361	345	**1.46**
AAPL	10	297	608	**1.36**
QQQ	15	214	253	**1.02**
SPY	15	178	226	**0.89**
TSLA	10	468	414	**0.89**

The deployed architecture wins on Winkler in **7 of 9 symbols**. The two exceptions are SPY and TSLA — the two deepest xstock-backed perp markets, and the symbols where the Cong et al. R^2 = 0.839 has the most empirical bite. For these names the tokenised perp tracks the NMS open competitively, and a non-parametric empirical-quantile residual

band on the perp is sufficient. The deployed architecture's edge widens as perp liquidity thins: for the long tail (HOOD, GOOGL, NVDA, MSTR, GLD, AAPL) the Winkler ratio is 1.36–$3.32\times$.

The pattern localises the advantage where the consumer's risk-management need is largest — on the thin-liquidity collateral that dominates the named lending-collateral universe (TSLAx aside, every other Kamino-onboarded xStock falls in the long tail above), not on the two deepest perps, where a simpler oracle is genuinely competitive. The defensible claim is: under a non-parametric tokenised-tracking baseline at its most-informed closed-window snapshot, the deployed architecture is the materially-tighter choice on 7 of 9 symbols and tied-to-modestly-behind on the other two. The architecture's separation of concerns admits a point-estimate swap — a tokenised-side conditional-mean import — without changing the conformal band scaffolding (Appendix F); the SPY/TSLA gap is a forecaster-swap surface, not a band-architecture failure.

D.7.6 Sample-size caveat and what is and is not powered

The bake-off panel is $n = 117$ weekend-symbol observations — about 7% of the $n = 1,730$ panel that backs §6 and D.1–D.6. At this sample size, the Kupiec / Christoffersen tests have limited power; under both the deployed architecture and the baseline, Kupiec p_{uc} at $\tau = 0.95$ accepts a wide band of realised-coverage values around nominal. **The powered signal in this comparison is the paired Winkler score** — a per-observation diagnostic that is much more powerful at small n than the binomial coverage tests. The pooled coverage results (Kupiec $p_{uc} = 0.949$ for M6 LWC, 0.742 for the baseline) are consistent with both methods being well-calibrated on this slice; they do not by themselves discriminate between architectures. The width-and-Winkler results do. We treat the §6 / D.1–D.6 panel as the powered backtest of the calibration claim itself, and D.7 as a (smaller-n) head-to-head against the tokenised-tracking competitor archetype that complements the D.1 (underlying-frozen) and §6.2 (GARCH-t) baselines on a different competitor axis.

Source script: `scripts/run_v1b_tokenised_tracking_baseline.py`; full panel artefact: `data/processed/v1b_tokenised_tracking_baseline.parquet`; tables: `reports/tables/paper1_c2_tokenised_tracking_baseline_summary.csv`, `reports/tables/paper1_c2_tokenised_tracking_baseline_per_symbol.csv`; full write-up: `reports/v1b_tokenised_tracking_baseline.md`.

D.7.7 Matched calibration history — isolating architecture from history length

The D.7.3 head-to-head gives the deployed band a decisive structural advantage the reader should see named: it is calibrated on the full 2014–2026 artefact, while the tokenised baseline can only calibrate over the post-2025-12 live perp tape (≤ 19 weekends per symbol). To separate the architecture from calibration-history length, we re-run both methods walk-forward on the *same* post-2025-12 window — refitting the deployed architecture (regime $\rightarrow$ per-symbol $\hat{\sigma} \rightarrow$ per-regime conformal quantile $\rightarrow \hat{\sigma}\cdot$`fri_close` band) on only the calibration data available up to each evaluation weekend, exactly as the baseline does, on the common $n = 90$ cells (`scripts/run_v1b_tokenized_matched_history.py`; tables `reports/tables/v1b_tokenized_matched_history_{summary,deltas,per_symbol,regime_floor}.csv`). Two concessions keep the comparison apples-to-apples and are themselves findings: the OOS $c(\tau)$ / $\delta(\tau)$ conservatism layer is dropped (it needs a train/OOS split the short window cannot afford, and the baseline has no such layer), and the $\hat{\sigma}$ warm-up is relaxed from ≥ 8 to ≥ 4 past observations (at ≥ 8 the short window strands $24/117$ cells).

Deployed-vs-baseline deltas at $\tau = 0.95$ (`mon_open` snapshot, block-bootstrap 95% CI):

calibration history	width reduction	Winkler improvement [95% CI]
frozen M6 (12 yr, re-scored on common $n = 90$)	43%	37.5% $[16.3, 51.8]$
matched M6, regime (≈ 4 mo)	26.7% $[14.9, 36.4]$	19.9% $[-1.5, 35.6]$
matched M6, pooled (≈ 4 mo)	29.6% $[18.4, 38.7]$	20.4% $[-2.0, 37.1]$

When calibration history is equalised, the width edge roughly halves ($43 \rightarrow 27$–30%) and the **Winkler edge collapses from ≈ 38–46% to $\approx 20\%$ with a 95% CI that spans zero** — Winkler is the coverage-fair arbiter, and it no longer distinguishes the architectures. The matched refit buys its remaining narrowness partly with lost coverage: it undercovers at 0.91 vs frozen-M6's 0.96 and the baseline's 0.94. At $\tau = 0.99$ the edge *reverses* (matched-regime Winkler -18%): the per-regime conformal quantile is 100% saturated at every walk-forward step in all three regimes (it needs ≥ 99 obs/cell; the largest cell reaches 63), so the tail band is systematically too narrow and the tail claim is simply not identifiable on the equalised window. Per-symbol, the deployment "wins seven of nine" falls to **four of nine**: matched-M6 beats the baseline only on the thin/volatile perps — GOOGL ($2.7\times$), HOOD ($3.1\times$), NVDA ($1.5\times$),

MSTR ($1.1\times$) — and loses on the deep-liquid core — SPY ($0.54\times$), QQQ ($0.61\times$), GLD ($0.68\times$), TSLA ($0.88\times$), AAPL ($0.95\times$) — where the perp tracks so tightly that a few-months-calibrated conformal band is wider than the baseline's empirical quantile.

Verdict for §6.2. The architecture advantage is real but much smaller than the deployment figure and confounded by coverage: roughly half to two-thirds of the $\tau = 0.95$ edge is the 12-year calibration record, not the architecture. The surviving architectural edge lives in the thin/volatile names — exactly where a lending consumer's closed-market risk concentrates — while the naive tracker is competitive-to-better on the liquid core. This is the empirical basis for the token-anchored hybrid named as future work (§9), and for the "data-hungry by construction" limitation (§8).

D.7.8 Jupiter-mid (v5/tape) cross-check — directional confirmation on the Solana-DEX surface

The kraken_futures xstock-backed perp is the deepest tokenised-equity surface but is not the on-chain SPL token a Solana-DEX consumer actually reads. The `soothsayer_v5/tape` jup_mid series (Jupiter's quoted mid for eight Backed xStocks) is the directly Solana-DEX-relevant signal; overlap with the deployed LWC artefact is one weekend (2026-04-24), extended two weekends forward via a forward-tape-extended $\hat{\sigma}$ lookup that re-uses the sidecar constants without re-fitting. Sample: 3 weekends × 8 symbols = $n = 24$. The baseline halfwidth is **borrowed verbatim** from the primary mon_open snapshot (no re-fit on the cross-check sample). The directional Winkler result reproduces — ratios baseline / M6 LWC at $\tau \in \{0.68, 0.85, 0.95, 0.99\}$ are $\{1.65, 2.09, 2.41, 2.11\}\times$. The ratio at $\tau = 0.95$ ($2.41\times$) is *wider* than the primary kraken_futures panel ($1.84\times$), consistent with Jupiter mid on Solana DEXs having a thinner, more dispersion-prone book than the deeper kraken_futures perp. Kupiec / Christoffersen are degenerate at $n = 24$ and we report only the Winkler / halfwidth axis. The cross-check is *confirmatory*, not powered. Source: `scripts/run_v1b_tokenised_tracking_v5_xcheck.py`; table `reports/tables/paper1_c2_tokenised_tracking_v5_xcheck.csv`.

E The serving layer

The deployment summary is in §4.9 — Python executable specification, Rust serving stack, Anchor on-chain program, byte-for-byte parity at 180/180 verification cases. This appendix is the engineering detail: the audience contract, the crate stack, the on-chain program and wire format, the parity harness mechanics, the integrator dependency path, and a worked $\tau \to$ reserve example.

E.1 Split-by-function architecture

The serving stack is a contract between three audiences: the **researcher** iterating against the Python OOS panel, the **protocol integrator** building against on-chain `PriceUpdate` accounts, and the **third-party auditor** verifying a served band against the published artefact. Byte-for-byte parity (E.5) makes the first two identical at the numeric level; the receipt fields and public artefacts make the third executable from public data alone. This trio is the operational instantiation of P_1 (auditability, §3.4).

The implementation separates *training* (Python) from *serving* (Rust + Solana). The Python `Oracle` (`src/soothsayer/oracle.py`) produces the per-Friday lookup artefact and is the executable reference for both the M5 Mondrian conformal-lookup algorithm (`Oracle.fair_value`) and M6's locally-weighted variant (`Oracle.fair_value_lwc`). No model logic is duplicated: the training pipeline materialises the per-Friday rows (including $\hat{\sigma}_s(t)$ per symbol pre-Friday) and the deployment scalars once offline; the serving stack reads them. The schedules are hardcoded in `oracle.py` and mirrored into both the JSON sidecar and `crates/soothsayer-oracle/src/config.rs` (M5 reference path + M6 LWC) for the Rust serving binary; Table A.2 (Appendix A) tabulates the schedule values.

Latency is a non-claim of the design (§3.5): serving is a five-line lookup against the materialised artefact, and nothing in the coverage-inversion mechanism requires a live forecast computation.

E.2 The Rust crate stack

Crate	Purpose	Lines	Tests
`soothsayer-oracle`	Library: loads the Mondrian artefact, exposes `Oracle::fair_value` with byte-exact parity to Python	~330	4/4 unit + parity harness
`soothsayer-publisher`	CLI binary: serves single reads, prepares on-chain publish payloads	~440	6/6 unit
`soothsayer-consumer`	`no_std`, minimal-dep SDK: decodes the on-chain `PriceUpdate` PDA into a typed `PriceBand`	~330	5/5 unit
`soothsayer-demo-kamino`	Reference Kamino-style consumer integration	~400	included in integration tests
`soothsayer-oracle-program` (Anchor)	On-chain program: `initialize`, `publish`, `set_paused`, `rotate_signer_set`	~400	scaffolded

The consumer SDK is `no_std` so it can be vendored into a Solana program account-decoder without the Rust standard library.

E.3 On-chain program

`programs/soothsayer-oracle-program/` is a vanilla Anchor program with three PDA account types (`PublisherConfig`, `SignerSet`, per-symbol `PriceUpdate`) and four instructions. The `publish` instruction validates signer membership, enforces the cadence rate-limit, and checks five on-chain invariants before persisting: band ordering ($L \le \hat{P} \le U$), coverage values $\le 10,000$ bps, exponent $\in [-12, 0]$, signer membership, and time since last publish $\ge$ `min_publish_interval_secs`. Every instruction emits an Anchor event so off-chain indexers can trace publish history without re-reading every account.

E.4 Wire-format design

The `PriceUpdate` account layout (128 bytes data + 8-byte Anchor discriminator) is informed by the Pyth-Network September-2021 BTC flash-crash post-mortem. Three design rules avoid the failure modes that incident exposed.

```
PriceUpdate:
  version              u8         schema version, currently 1
  regime_code          u8         0=normal, 1=long_weekend, 2=high_vol
  forecaster_code      u8         0=F1_emp_regime, 1=F0_stale (both legacy v0);
                                   2=mondrian (M5; live on-chain);
                                   3=lwc (M6; live in Rust, on-chain slot
                                   reserved pending next publisher release)
  exponent             i8
  target_coverage_bps  u16        publisher-set target coverage, integer bp
  claimed_served_bps   u16        served quantile, integer bp
  buffer_applied_bps   u16
  symbol               [u8; 16]   ASCII null-padded
  point, lower, upper, fri_close   i64        absolute fixed-point at ‘exponent‘
  fri_ts, publish_ts, publish_slot, signer, signer_epoch    timestamps + identity
```

The three rules: (1) **Absolute prices, not deltas** — a consumer reading `lower` gets a real price directly. (2) **Single shared exponent** for all price fields, so cross-field comparisons are exact integer comparisons. (3) **Integer basis points, never floats** for coverage and buffer — exact and deterministic across Rust, Anchor IDL, TypeScript clients, and Python. Wire-format sizes are locked by unit tests in `crates/soothsayer-publisher/src/payload.rs`.

On the receipt side, the `calibration_buffer_applied` field carries $\delta(\tau)$ on the receipt (identically zero under the deployed schedule, §4.5); the diagnostics carry $c(\tau)$ and q_{eff}.

E.5 Byte-for-byte parity verification

§4.9 states the parity headline; the mechanics are as follows. `scripts/verify_rust_oracle.py` runs Python `Oracle.fair_value` (M5) / `Oracle.fair_value_lwc` (M6) and the Rust `soothsayer fair-value --forecaster {mondrian, lwc}` CLI on the same `(symbol, fri_ts, target_coverage)` triples and asserts byte-exact agreement on numeric output (point, lower, upper, sharpness_bps, claimed_served, calibration_buffer_applied) plus exact agreement on string fields (regime, forecaster_used). Each forecaster's panel covers 90 cases (25 random pairs from the artefact + edge cases for SPY, GLD, TLT, MSTR, HOOD at three targets each). **180/180 pass** as of the most recent re-run — 90/90 on the M5 path and 90/90 on the M6 LWC path. The Rust crate serves both forecasters; `forecaster_code = 3` (`FORECASTER_LWC`) is encoded into new publishes from `crates/soothsayer-publisher` and decoded into the `Forecaster::Lwc` variant by the no_std `soothsayer-consumer` crate.

The parity harness is invoked after every methodology change: each migration is applied to Python first, then ported to Rust, and the harness re-runs before the change is considered shipped. It is the contract that allows a consumer to verify a served `PricePoint` against the published artefact without trusting either implementation in isolation.

Wire format invariance. The on-chain `PriceUpdate` Borsh layout is unchanged across all methodology migrations to date. The only on-wire change at each step is relabelling reserved `forecaster_code` slots: code 2 → `FORECASTER_MONDRIAN` (M5; live), code 3 → `FORECASTER_LWC` (M6; live in the Rust serving stack as of the 180/180 parity milestone above; on-chain enablement gated on the next publisher release). Older accounts decode cleanly under newer consumers; M6 accounts decode cleanly under any consumer that reads the `forecaster_code` byte (`soothsayer-consumer` exposes the typed `Forecaster::Lwc` variant). Downstream consumers branching on `forecaster_code` need only add code-2 and code-3 cases.

E.6 Reproduction and integrator path

Build, artefact-freeze, parity-verification, and devnet-deployment commands are consolidated in Appendix A (§A.5) and are not duplicated here.

A lending-protocol integrator depends on `soothsayer-consumer` (the `no_std` decoder) and reads the `PriceUpdate` PDA at the symbol's derived address; the demo Kamino-style consumer at `crates/soothsayer-demo-kamino/` shows the full integration in ~400 lines, including LTV decision logic that uses the band's lower bound as the collateral valuation input — the third audience of E.1 reading the same wire bytes the on-chain program emits.

E.7 Practitioner integration — a worked $\tau \to$ reserve example

To make the consumer contract concrete we walk one illustrative mapping from served coverage to a reserve decision. It is an *example of the mechanics*, not a recommended policy: choosing τ, the LTV ladder, and reserves to minimise expected protocol loss under an explicit cost model is the decision-theoretic problem held out of scope (§3.5).

Consider a market holding SPYx as collateral with a liquidation-LTV threshold θ and a position at current LTV $\ell < \theta$. Its *adverse-move buffer* — the fractional collateral drawdown that exhausts the position's headroom — is $b = 1 - \ell/\theta$. The protocol reads the served lower bound $L_\tau = \hat{P}\,(1 - d_\tau)$ and treats the fractional band drawdown $d_\tau = q_{\text{eff}}(\tau)\,\hat{\sigma}_s$ as the closed-market collateral shock it must survive with probability τ.

- τ **selection.** Choosing τ sets the per-name closed-market breach budget to $1 - \tau$: at $\tau = 0.95$ the served lower bound is breached on $\sim 5\%$ of windows, at $\tau = 0.99$ on $\sim 1\%$. A protocol picks τ so its tolerated per-name bad-debt frequency matches $1 - \tau$.
- **Reserve headroom.** A position survives the τ-band shock iff $b \geq d_\tau$. For SPYx at $\tau = 0.95$ the served half-width is ≈ 178 bps ($d_\tau \approx 1.78\%$; Appendix D §D.7.5); against the production origination-to-liquidation buffer of $\sim 2.7\%$ on SPYx/QQQx, a typical position clears the $\tau = 0.95$ band shock with $\sim 0.9\%$ to spare, but a position opened near the liquidation threshold ($b < 1.78\%$) does not and should be pre-emptively de-risked or reserved against. This is the **narrow-buffer** reserve class where the band has decision-flipping headroom; wide-buffer names (14–25% buffers) are only bound by the band at $\tau = 0.99$ or under the joint tail.
- **Portfolio reserve.** For a book of m correlated names the single-name d_τ does not aggregate independently: the protocol sizes its reserve against the joint breach-count k_w distribution and applies the circuit-breaker threshold k^* — both measured and published in §8, whose numbers this example inherits by cross-reference rather than restating.

Every quantity above is on the served receipt ($\hat{P}$, $d_\tau = q_{\text{eff}}\hat{\sigma}_s$, τ) plus the public k_w CDF (§8); no incumbent oracle exposes the inputs this mapping needs. The full policy optimisation is out of scope (§3.5).

F Oracle comparator detail and extended diagnostics

This appendix has two halves. The first carries the full methodology behind Table T1's measured findings (§1.2): the Pyth availability and multiplier-ladder measurements, the Chainlink v11 synthetic-marker classification, and the per-venue schema and wire-field evidence summarised in §2. The second carries the extended diagnostics behind §8: the cluster topology of the full-distribution residual, the CUSUM drift bank, the unmeasured execution-side effects, the scope boundaries, and a full-distribution architectural variant that is characterised but not deployed.

F.1 Scope and attribution

No incumbent oracle surface emits a calibrated coverage band, so a coverage-vs-sharpness comparison against them is not well-defined and is not reported (§2). The measurements below therefore characterise something narrower and precisely stated: what realised coverage a *consumer* receives if they read an incumbent's published fields as a probability statement. Where a multiplier or a scaling is required to reach a target coverage level, that calibration is the consumer's, not the venue's — no venue publishes one, and no venue's stated claim is contradicted by any measurement here.

One exclusion for integrity: the source report behind the Pyth measurements (`reports/v1b_pyth_comparison.md`) also contains a band-width comparison against a v1-era Soothsayer surface. That comparison predates the deployed M6 architecture and is excluded here; only the Pyth-side measurements port.

F.2 Pyth dispersion field — availability and the multiplier ladder

Method. For each (symbol, fri_ts) in the 2024+ subset of the OOS panel (1,200 symbol × weekend pairs across 120 weekends), pull Pyth's (price, conf) via the Hermes Benchmarks API at the nearest publish to Friday 15:55 ET, retrying up to 2 hours back if the queried timestamp falls outside Pyth's publish cadence. For $k \in \{1, 1.96, 3, 5, 10, 25, 50, 100, 250, 500, 1000\}$, define the implicit band [price $- k \cdot$ conf, price $+ k \cdot$ conf] and check whether the realised Monday open lies inside.

Panel restriction. The panel is restricted to 2024+ because Pyth's regular-hours US-equity feeds did not exist before then; queries against 2023 timestamps return empty consistently across all tested tickers. The comparison therefore covers the 120 weekends of the 2024+ window within the 2023+ OOS slice. Pyth's documentation describes `conf` as a publisher-dispersion diagnostic, not a probability-of-coverage claim [Network, 2022], so any multiplier k found below is a *consumer-supplied calibration*, not a Pyth-published one.

Per-symbol availability.

symbol	Pyth available rate
SPY	0.692
QQQ	0.650
TLT	0.592
TSLA	0.250
MSTR	0.017
GLD	0.008
AAPL	0.000
GOOGL	0.000
HOOD	0.000
NVDA	0.000

Pooled availability across 2024+ is **22.1%** (265 of 1,200 attempts had Pyth data within ±2h of Friday close). The coverage analysis below uses the 265-obs subset; on this subset the symbol mix is SPY-, QQQ-, TLT- and TSLA-heavy — large-cap, low-realised-volatility names by composition, which biases the sample toward easier weekends than the pooled OOS slice.

Pooled realised coverage at increasing k.

k	n	realised	mean half-width (bps)
1.00	265	0.049	5.6
1.96	265	0.102	11.0
3.00	265	0.155	16.8
5.00	265	0.242	27.9
10.00	265	0.434	55.9
25.00	265	0.800	139.7
50.00	**265**	**0.951**	**279.5**
100.00	265	0.992	559.0
250.00	265	1.000	1397.4
500.00	265	1.000	2794.9
1000.00	265	1.000	5589.7

A consumer reading `conf` as a Gaussian standard deviation ($k = 1.96$, the textbook 95% wrap) realises **10.2%** coverage on the available subset; tripling the field ($k = 3$) recovers only 15.5%. The `conf` field is tight relative to weekend-gap dispersion because it measures publisher dispersion at venue close, which is not the object a closed-market coverage statement needs. The $k = 50$ that realises 95% on this sample is fit in hindsight on the evaluation data: no ex-ante-choosable multiplier on the published dispersion yields a known coverage level, and a consumer who back-fits k on their own historical sample has constructed their own calibration — with no audit trail and no per-symbol stratification — rather than read one off the wire. This is the k-ladder finding carried in Table T1.

Caveats. (i) The benchmark uses Pyth's regular-hours feeds (`Equity.US.{TICKER}/USD`); Pyth also publishes overnight (`.ON`) and pre/post-market (`.PRE`/`.POST`) feeds not probed here. (ii) AAPL, GOOGL, HOOD, NVDA show 0% availability at the 15:55 ET probe despite valid feed IDs — consistent with session-restricted publishing at the probe time or feed launches after the window opened. (iii) The ladder values are point estimates on a 265-obs sample without bootstrap CIs. Raw observations: `data/processed/pyth_benchmark_oos.parquet`; per-(k, scope) breakdown: `reports/tables/pyth_coverage_by_k.csv`; runner: `scripts/pyth_benchmark_comparison.py`.

F.3 Chainlink Data Streams v11 — synthetic weekend markers

v11 is the active 24/5 US-equity report schema (schema id `0x000b`, 14-field 448-byte ABI-encoded payload) [Labs, 2026]; its `market_status` enum encodes the underlying venue session, with `market_status = 5` covering weekends. Public documentation does not state what `bid`/`ask`/`mid` carry during `market_status` $\neq$ `2` (regular session); the schema still publishes a price on weekends, and whether that price is real or placeholder-derived is an empirical question the wire answers.

Panel and classifier. An 87-observation weekend panel (2026-02-06 $\rightarrow$ 2026-04-17; `reports/v1b_chainlink_comparison.md`) establishes that v11 publishes a price during weekend windows but no coverage band — the wire `bid`/`ask` are the closest fields. A synthetic-marker classifier (v2, scan of 2026-04-26; `reports/v11_cadence_verification.md`) decoded 26 v11 weekend reports at `market_status = 5` — 20 on the four mapped xStocks (SPYx $n = 6$, QQQx $n = 6$, TSLAx $n = 7$, NVDAx $n = 1$) plus 6 on unmapped tickers — and labelled each against a six-class taxonomy distinguishing real quotes from placeholder bookends.

Per-symbol findings.

- **SPYx — pure placeholder.** Both bid and ask carry the synthetic `.01`-suffix marker in 100% of weekend samples (bid 21.01, ask 715.01; spread 18,858 bps). The `mid` field (368.0100) is the arithmetic midpoint of the synthetic bookends, not a market mid; `last_traded_price` (713.96) is plausibly the real Friday close.
- **QQQx, TSLAx — bid-synthetic.** The bid carries the `.01` marker in 100% of weekend samples (QQQx 656.01, TSLAx 372.01) while the ask is not so marked; spreads 117–329 bps. Real-market bids land on `.01` at roughly a 1-in-100 base rate; a 100% incidence is a strong synthetic signal.
- **NVDAx — mixed.** The single weekend sample ($n = 1$) classifies as real (bid 208.07, ask 208.14, spread 3.4 bps). Single-sample evidence; it needs replication before generalising, and it leaves open whether the synthetic-marker pattern is per-feed rather than universal.
- **Sessions 1/3/4** (pre-market, post-market, overnight) had no samples in the scan window and are unmeasured.

Two consequences for a consumer. First, the spread is a misleading signal on weekends: the pure-placeholder pattern produces an 18,858-bps spread while the bid-synthetic pattern can fall under 200 bps, so a spread-threshold filter misses the synthetic-low mechanism — classification requires the marker check. Second, the point estimate survives

where the interval does not: the measured pooled weekend point-estimate bias is -8.77 bps (undetectable at panel size; `reports/v1_chainlink_bias.md`) — the synthetic-bid / real-mid combination yields a usable point and no usable interval. These are the synthetic-marker findings carried in Table T1, with sample bounds as stated (26 decoded reports, 20 on the four mapped xStocks, `market_status = 5`, Feb–Apr 2026).

F.4 Per-venue schema and wire-field evidence

§2 states the categorical claim — no surveyed incumbent surface publishes a calibration claim on the served value; this section catalogues the per-venue field evidence behind it.

Chainlink Data Streams ships three co-existing RWA schemas as of 2026-Q2: v8 RWA Standard [Labs, 2025b], v10 Tokenized Asset [Labs, 2025a], and v11 RWA Advanced [Labs, 2026]. v8 is the load-bearing path for tokenized-equity lending on Solana: Kamino Finance consumes every xStock collateral asset through Kamino Scope's `OracleType::ChainlinkRWA` adaptor [Finance, 2025b,a], routed through the v8 dispatcher, with two parallel feeds per xStock registered in Scope's mainnet config [Finance, 2026] — an `Open` mode that refuses updates outside US trading hours and an `AllUpdates` mode clamped to ±500 bps of the Open feed's Friday-close value. On the wire, v8 is mid-only (no `bid`/`ask`/confidence) with `marketStatus` as the only off-hours regime signal. v10 backs the Backed↔Solana xBridge but is not consumed by the canonical Solana lending market; it carries no `bid`/`ask`/confidence field and serves a 24/7 `tokenizedPrice` whose CEX-aggregation methodology is qualitatively described. v11's weekend `bid` carries the synthetic `.01`-suffix at 100% incidence on three of four mapped xStocks (F.3).

Pyth [Network, 2021, 2022] aggregates first-party publisher submissions: each permissioned publisher submits a price and confidence interval, and the aggregate rule assigns three votes per publisher (price, price ± confidence) before taking the median. The published `(price, conf)` reports low publisher disagreement at venue close; documentation asserts ~95% coverage at the *publisher* level, which is per-publisher self-attestation, not a property of the aggregate a protocol reads. Pyth Pro layers subscriber-customisable cadence and the Blue Ocean ATS overnight integration [Network, 2026, 2025], covering 8:00 PM – 4:00 AM ET Sun–Thu — the canonical xStock weekend remains uncovered.

CF Benchmarks (Kraken-owned, FCA + BMR-supervised) launched the xStocks Indices family on 2026-05-07 [Benchmarks, 2026b,a]: regulated per-second scalars per xStock (Price Return + Total Return variants) powering Kraken's tokenized-equity perpetuals. The BMR Article 27 Benchmark Statement contains no language about realised coverage; no confidence, band, or coverage field is on the wire.

SEDA publishes session-aware single-asset feeds and a USA500 composite (60% equities / 25% crypto / 15% commodities) consumed by Hyperliquid HIP-3 markets via Dreamcash, Injective, and Perps.Fun [Protocol, 2026], governed by a four-pillar continuity rule (session tags, cost-of-carry futures-to-spot, self-referencing EMA, staleness fallback); the output is a single scalar plus session-flag metadata.

Others. Switchboard generalises aggregation into permissionless queues with shallow tokenized-equity deployment [Switchboard, 2023]; RedStone Live blends institutional feeds with perpetual-market data into a 24/7 scalar [RedStone, 2026]; optimistic oracles UMA [Project, 2020] and Tellor [Tellor, 2023] guarantee *eventual* correctness under economic security but define no distributional statement conditional on time-of-week or realised volatility.

F.5 External-evidence epistemic status

Two external-evidence inputs to the §1/§2 framing warrant explicit disclosure.

Cong et al. (2025) panel [Cong et al., 2025]. The empirical premise — that tokenized-equity prices deviate from underlying closes with documented frequency during closed-market windows, that the wedge is structural rather than transitional, and that off-hour tokenized returns carry a conditional-mean signal on the underlying open — leans on Cong, Landsman, Rabetti, Zhang and Zhao. Their sample is four months (July–October 2025), 100 tokenized stocks, $420M aggregate market cap concentrated in Backed Finance and Ondo Global Markets, and a late-2025 volatility regime. The deviation-frequency table cited most directly (Table 4, p.32) is for TSLAx alone; the panel-level passthrough regression (Table 10, p.40: $\lambda = 0.903$, adjusted $R^2 = 0.839$) pools across the 100 names without per-symbol stability evidence. We do not assume any specific Cong-cited number carries forward in distributional shape; the citations establish the empirical context for the §1 motivation, and the deployed primitive's coverage claim is validated independently on the twelve-year ten-symbol panel of public-data underliers (§6).

Kamino's per-reserve `Open` / `AllUpdates` mapping. Scope's dual-feed architecture (F.4) offers two serving modes per xStock; we do not pin which mode any given xStock lending reserve currently subscribes to. The per-reserve selection is governance-set in the mainnet config [Finance, 2026], but the live `kamino-lending` reserve PDA mapping is governance-mutable and was not decoded end-to-end. The claim as stated — neither mode publishes a calibrated

coverage band, and the choice between them is a governance parameter — is stable across reconfigurations; if the on-chain decoding completes, the claim sharpens to the single archetype actually selected without weakening the categorical contribution. This bound is also stated in Table T1's caption (§1.2), whose rows 1–2 describe the two available modes, not a per-reserve assignment.

F.6 Cluster topology of the full-distribution residual

§8 states the headline: the served band is per-anchor calibrated but not full-distribution calibrated, and the residual is within-weekend cross-sectional co-breach. This section carries the localisation detail.

Cluster structure. The residual is a within-weekend cluster structure separating equities (8 symbols, positively correlated with the weekend equity factor) from safe-haven assets (GLD, TLT, with negative or near-zero correlation to the equity factor). Within the equity cluster $\hat{\rho}_{\text{within}} = 0.477$ — *higher* than the panel-wide mean intra-weekend pairwise correlation $\bar{\rho}_{\text{intra}} = 0.41$, because the cross-cluster correlations dilute it. Within the safe-haven cluster $\hat{\rho}_{\text{within}} = -0.016$ — there is no common-mode to remove.

Why a naive partial-out fails. A non-parametric partial-out using either the within-weekend median or mean reduces $\bar{\rho}_{\text{intra}}$ from 0.41 to $\{0.12, 0.06\}$ but breaks per-symbol Kupiec on the safe-haven cluster (GLD violation rate 14.5%, TLT 9.2% under the median variant; $5/10$ per-symbol pass under the mean variant) and rejects DQ at $\tau \in \{0.68, 0.85\}$. The mechanism is structural: subtracting an equity-dominated weekend median from a negatively-correlated asset creates fresh violations on equity-rally weekends and pulls residuals toward the band centre on selloff weekends. Any operation that treats the panel as exchangeable inherits this failure mode; a working architecture-level fix must respect the cluster topology (F.10).

Two views of the same structure. The over-dispersion of the homogeneous t-copula (§8; emp/t variance 0.63 at $\tau = 0.95$) and the cluster failure of the naive partial-out are two views of the same equity-vs-safe-haven structure: a single Pearson $\hat{R}$ averages high equity-equity correlations against negative equity-safe-haven correlations; a single weekend median centres on the equity cluster's mean. Either rendering of the panel as homogeneous misses the same topology. (Source: A1, A1.5, A3 — `paper1_a1_*.csv`, `paper1_a1_5_*.csv`, `paper1_a3_*.csv`.)

Localising the lag-1 alternative. The AR(1) signal lives in the *cross-sectional within-weekend* ordering ($\hat{\rho}_{\text{cross}} = 0.354$, $p < 10^{-100}$), not the temporal-within-symbol ordering ($\hat{\rho}_{\text{time}} = -0.032$, $p = 0.18$). $\hat{\sigma}$ standardisation operates *within* a symbol's history, not *across* symbols within a weekend, so the deployed architecture is structurally unable to reach this residual. A vol-tertile sub-split of the `normal` regime (5 cells) leaves Berkowitz LR essentially unchanged ($170 \to 172$) while widening the $\tau = 0.95$ band by 9% — finer regime granularity is not the lever. Within-bin exchangeability separately holds (permutation test on lag-1 score autocorrelation within each regime × symbol bin: $1/30$ nominally rejecting at uncorrected $\alpha = 0.05$, $0/30$ after Benjamini–Hochberg correction; Appendix B). The cross-sectional dependence is *across* bins, not within.

Per-symbol Berkowitz localisation. The residual symbols rejecting Berkowitz at $\alpha = 0.01$ are TLT (LR $= 16.7$) and TSLA (LR $= 13.5$) — both trace to cross-sectional common-mode rather than per-symbol scale. The joint-tail distribution of §8 turns the Berkowitz rejection into a measurable, reportable shape ($2.34\times$ variance overdispersion vs independence at $\tau = 0.95$; $P(k_w \geq 3) = 4.62\%$ vs binomial 1.15%, t-copula 8.2%), and the k^* threshold gives the consumer an operational handle that holds up against the joint-volatility model a CCAR-style risk team would run, not just against independence.

F.7 τ-stratified CUSUM drift bank

The CUSUM drift monitor — calibration, detection power, and its SVB (2023-03-10) and BoJ (2024-08-02) firings — is documented in Appendix B §B.17, alongside the forward-tape harness whose weekly ingest path it re-uses; §8 and §9 cite it there. The residual it monitors is the cross-sectional co-breach structure of §F.6.

F.8 MEV and execution-aware coverage — not measured

A consumer transacting at a worse price than the band's mid (front-running, sandwich, block-builder reordering) experiences an *effective* coverage that may differ from the measured path coverage near band edges. Measuring this requires Solana mempool-or-bundle granularity plus a consumer-behaviour model; neither is in this paper's data, and no execution-aware coverage number is claimed. The data requirement is stated with the path-fitted variant in F.11.

F.9 Region and asset-class scope

Region. All ten tickers are US-listed; the closed-market window predicted is the US weekend or US overnight, and the factor switchboard privileges US session data. We make no claim about generalisation to tokenized-JP or tokenized-EU equities.

Asset class. The methodology requires a *continuous off-hours information set* — a public, free, sub-daily price signal correlated with the closed-market underlier. Tokenized US equities have this (E-mini futures + VIX), tokenized gold has it (gold futures + GVZ), tokenized US treasuries have it (10Y T-note futures + MOVE), and most major FX pairs have it; tokenized credit has it through CDS spreads. Out of scope: real estate (no continuous off-hours signal), illiquid commodities (futures discontinuous or thin), and instruments with primarily NAV-based pricing. `docs/methodology_scope.md` carries the per-class fit/no-fit table.

F.10 Full-distribution conformal variant — characterised, not deployed

Per §9: this variant is characterised, not deployed, and nothing the paper claims depends on it.

The pooled Berkowitz and DQ rejections at $\tau = 0.95$ (§8) are the deepest residual. The deployed architecture is *partially* conformal — per-regime Mondrian quantiles on per-symbol-$\hat{\sigma}$-standardised residuals with a near-identity $c(\tau)$ bump; a full-distribution variant replaces the four-anchor schedule with a single conformal correction over the residual CDF, delivering exchangeability-based finite-sample coverage at every τ simultaneously (Romano–Patterson–Candès CQR [Romano et al., 2019] is the closest match). Because the residual localises to the equity-vs-safe-haven cluster topology (F.6), the correction must be cluster-conditional: a cluster-internal partial-out within the equity cluster reduces $\hat{\rho}_{\text{within}}$ from 0.477 to 0.077, preserves per-symbol Kupiec 8/8 on the cluster, and tightens half-width -11.6% at $\tau = 0.95$. Composed with the path-fitted conformity score of F.11, the characterisation is τ-conditional: at $\tau \geq 0.95$ the cluster-conditional + path-fitted combination is empirically calibrated on the equity-cluster + CME-projected subset, moving pooled DQ at $\tau = 0.95$ from rejection to $p = 0.912$; at $\tau < 0.95$ endpoint scoring on the unpartialled residual remains dominant — path extrema introduce within-symbol temporal autocorrelation that DQ catches at body-of-distribution τ. The recommendation is bounded to the anchors where it is calibrated, not blanket.

F.11 Path-fitted conformity score and the earnings-night bake-off — technical detail

Path-fitted score. The path-fitted conformity score is implemented and unit-tested (`soothsayer.backtest.calibration.compute_score_lwc_path`; 7 unit tests, all pass), and an empirical first cut on the CME-projected subset ($n = 1,557$ OOS rows) confirms mechanical correctness. Continuous-consumption AMM use cases require this path-fitted variant rather than the endpoint score — the endpoint-not-path bound of §8. The binding empirical question — whether path-fitting closes the measured endpoint-vs-path gap on consumer-experienced path data — is answerable once the V5 forward-cursor tape (continuous capture since 2026-04-24) accumulates $N \geq 300$ path-coverage weekends. An MEV-conditional variant additionally requires Solana bundle-level granularity to reconstruct the price a consumer would have *executed at* near band-edge events (F.8).

Earnings-night bake-off. The `earnings_night` regime (§6.5) is the underlying-side confirmation of Cong et al.'s value-relevant-news mechanism; the sharper head-to-head is the calendar-conditioned band against the Cong-implied tokenized passthrough baseline (`point = tokenised_price_at_t`; band = ±k × historical residual), evaluated under matched-τ Kupiec / Christoffersen *on earnings nights specifically*, once the V5 forward-cursor tape accumulates enough post-launch earnings nights. That comparison is the empirical bridge from the present underlying-side result to a tokenized-side claim, and it is deliberately not asserted here.

G Appendix G — The lending-track instantiation: one-sided downside bands

The deployed architecture of §4 serves a two-sided band. A lending protocol holding tokenised equity as collateral is exposed in **one** direction: the collateral falling below the debt it secures. This appendix instantiates the §3 primitive with a one-sided downside score, quantifies what the two-sided band costs a consumer who only faces downside, and states what would be required before it replaces the two-sided path. It is characterised, not deployed — the same standing as the full-distribution variant of Appendix F.

G.1 G.1 Why the direction matters to the contract, not the primitive

The §3 contract is direction-agnostic. Target coverage τ is the consumer's input, the band is the output, and the receipt exposes the sufficient statistics for re-derivation. Nothing in that construction requires the served interval to be

symmetric about the point estimate; symmetry enters only through the conformity score of §4.2, which is an absolute value.

That choice has a consequence the consumer does not see on the wire. A two-sided band at τ places $(1-\tau)/2$ in each tail, so **its lower edge is a** $(1+\tau)/2$ **one-sided bound.** A protocol wiring the $\tau = 0.85$ band into a loan-to-value haircut is provisioning to 92.5%, and one wiring $\tau = 0.95$ is provisioning to 97.5%. The number they requested and the protection they received differ, and the gap is paid in collateral.

This is also where the paper meets the conformal-VaR literature it cites. Both Schmitt [Schmitt, 2026] and Zhong [Zhong, 2026] calibrate *one-sided* VaR; the two-sided choice here is ours, not the field's, and it is a modelling convenience rather than a property of the primitive.

G.2 G.2 Construction

Everything is held fixed against §4 except the score. Same factor-switchboard point estimator, same per-symbol $\hat{\sigma}_s(t)$, same Mondrian cells, same finite-sample rank $\lceil \tau(n+1) \rceil$, same 2023-01-01 split, and the same OOS-fit $c(\tau)$ bump — applied to *every* arm, so no arm is advantaged. Because the $c(\tau)$ grid begins at 1.0 it can only widen; it repairs under-coverage and cannot manufacture a capital saving.

$$s^{2s} = \frac{|P_{\text{Mon}} - \hat{P}|}{P_{\text{Fri}}\,\hat{\sigma}_s}, \qquad s^{1s} = \frac{\hat{P} - P_{\text{Mon}}}{P_{\text{Fri}}\,\hat{\sigma}_s}$$

s^{1s} is signed and positive when the realised price lands *below* the point estimate, so its upper τ-quantile is the downside buffer directly. The served lower edge is $\hat{P} - q_r(\tau)\,\hat{\sigma}_s\,P_{\text{Fri}}$ and no upper edge is published.

Two comparisons are reported, because they answer different questions. **Same-label** compares the one-sided band at τ against the two-sided band this paper deploys at τ — it prices the hidden conservatism. **Matched-protection** compares the one-sided band at τ against the two-sided band at $2\tau - 1$, both of which *claim* the same downside coverage — it asks whether a directly-fitted downside quantile is better than an absolute-value quantile at equal protection.

G.3 G.3 Same-label: what the two-sided band costs

Collateral buffer, measured as the distance from the point estimate to the served lower edge in basis points of P_{Fri}, on the 2023+ holdout:

τ	panel	two-sided (deployed)	one-sided	buffer freed
0.68	weekend	130.8	55.0	-57.9%
0.85	**weekend**	**213.6**	**140.7**	$\mathbf{-34.1\%}$
0.95	weekend	343.4	292.8	-14.7%
0.99	weekend	633.1	589.6	-6.9%
0.68	overnight	112.7	34.7	-69.2%
0.85	**overnight**	**178.9**	**105.7**	$\mathbf{-40.9\%}$
0.95	overnight	280.1	213.3	-23.8%
0.99	overnight	474.3	402.8	-15.1%

At the default deployment anchor $\tau = 0.85$ a consumer frees roughly a third of the collateral buffer **and** receives the protection level it requested rather than an unlabelled tighter one. The saving compresses as $\tau \to 1$ because the two tails converge in relative terms; it is largest exactly where the deployment default sits.

G.4 G.4 Matched-protection: the one-sided fit is better calibrated

At equal claimed downside coverage the capital comparison is close to a wash and splits by anchor — the one-sided fit is narrower in the body (overnight -21.4%, -8.2%, -2.7% at $\tau = 0.68, 0.85, 0.95$) and wider in the tail (weekend $+11.5\%$ at 0.95, $+17.0\%$ at 0.99). That ordering is expected rather than anomalous: the absolute-value score pools both tails and therefore carries roughly twice the effective sample for a tail quantile. Pooling helps where data is thin and costs where the symmetry assumption is wrong.

Calibration is not a wash, and it is the substantive result. Kupiec p on downside breaches at matched protection:

	$\tau = 0.68$	$\tau = 0.85$	$\tau = 0.95$	$\tau = 0.99$
one-sided, weekend	0.975	0.973	0.956	0.942
two-sided at $2\tau - 1$, weekend	0.287	0.569	0.214	0.276
one-sided, overnight	1.000	0.958	0.157	0.950
two-sided at $2\tau - 1$, overnight	**0.000**	**0.000**	**0.045**	0.950

Overnight, imposing a symmetric shape on the gap distribution **fails Kupiec at three of four anchors** and holds per-symbol on only 5 of 10 and 8 of 10 symbols at $\tau = 0.68$ and 0.85; it over-covers systematically (0.710 against a claimed 0.680). The one-sided fit passes at every anchor on both panels. The symmetry assumption is measurably wrong for overnight gaps, and the direction of the error is conservative — which is why it has been invisible.

So the one-sided instantiation is not a capital optimisation dressed as a methodology change. It is the correct calibration of the quantity a lending consumer acts on.

G.5 G.5 What is not established

This appendix reports a single experiment on the 2023+ holdout. It does **not** carry the evidence pack that supports the two-sided architecture in §6 and Appendices B–D: no leave-one-symbol-out generalisation, no nested temporal holdout, no forward-tape validation against a frozen artefact, no simulation study on synthetic ground truth, no block-bootstrap confidence intervals, no DQ or Berkowitz diagnostics, and no Rust parity. The two-sided path remains the validated deployment and the basis of every claim in the main text.

Three items would have to close before a one-sided band could replace it:

1. **The full validation battery**, re-run on the one-sided score. The machinery is score-parameterised, so this is largely re-execution rather than new construction.
2. **A new frozen artefact and its own forward tape.** This is the binding constraint and it is calendar-bound rather than effort-bound: a new freeze restarts forward accumulation at $N = 0$ and accrues one weekend per week. The two-sided freeze carries roughly a quarter of accumulated forward evidence that a switch would forfeit.
3. **A wire-format decision.** `PriceUpdate` currently carries a symmetric $(\text{point}, \text{half\_width})$; a one-sided band changes the consumer contract of §8. No integrator is live at the time of writing, so the migration cost is presently nil — a window that closes with the first integration.

G.6 G.6 Consumer guidance

Independent of deployment, G.3 changes what a protocol should be told. A consumer whose exposure is one-directional should read the deployed two-sided band at τ as a $(1 + \tau)/2$ downside bound and size collateral accordingly, or request the anchor $2\tau - 1$ to obtain τ downside protection from the two-sided path. The integration guidance of §8 states the symmetric interval without stating that consequence; that is a documentation gap in the current release, not a defect in the served value, and G.3 quantifies what it costs.

References

A. Aich, A. B. Aich, and D. C. Jain. Temporal conformal prediction (TCP): a distribution-free statistical and machine learning framework for adaptive risk forecasting. arXiv:2507.05470v6 [stat.ML] (v6 dated 22 January 2026), 2025. URL `https://arxiv.org/abs/2507.05470`.

S. Allen, J. Koh, J. Segers, and J. Ziegel. Tail calibration of probabilistic forecasts. Journal of the American Statistical Association 120(552), 2796–2808 (published online December 22, 2025), 2025. URL `https://doi.org/10.1080/01621459.2025.2506194;arXiv:2407.03167`.

A. N. Angelopoulos, E. J. Candès, and R. J. Tibshirani. Conformal PID control for time series prediction. NeurIPS 2023; arXiv:2307.16895v1 [cs.LG], 2023. URL `https://arxiv.org/abs/2307.16895`.

G. Angeris and T. Chitra. Improved price oracles: constant function market makers. Proceedings of the 2nd ACM Conference on Advances in Financial Technologies (AFT '20); arXiv:2003.10001, 2020. URL https://arxiv.org/abs/2003.10001.

R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. Annals of Statistics 51(2), 816–845, 2023. URL https://doi.org/10.1214/23-AOS2276;arXiv:2202.13415.

M. J. Barclay and T. Hendershott. Price discovery and trading after hours. Review of Financial Studies 16(4), 1041–1073, 2003. URL https://doi.org/10.1093/rfs/hhg030.

M. J. Barclay and T. Hendershott. Liquidity externalities and adverse selection: evidence from trading after hours. Journal of Finance 59(2), 681–710, 2004. URL https://doi.org/10.1111/j.1540-6261.2004.00646.x.

CF Benchmarks. Token market price benchmarks series — methodology guide (v1.0, 02 feb 2026) and benchmark statement (BMR article 27 disclosure). CF Benchmarks documentation portal, 2026a. URL http://docs.cfbenchmarks.com/Token%20Market%20Price%20Benchmarks%20Series%20-%20Methodology%20Guide.pdf;https://docs.cfbenchmarks.com/Token%20Market%20Price%20Benchmarks%20Series%20-%20Benchmark%20Statement.pdf.

CF Benchmarks. CF benchmarks xStocks product suite is live with regulated indices and a corporate-actions feed. CF Benchmarks blog (launch announcement, 2026-05-07), 2026b. URL https://cfbenchmarks.com/blog/cf-benchmarks-xstocks-product-suite-is-live-with-regulated-indices-and-a-corporate-actions-feed.

J. Berkowitz. Testing density forecasts, with applications to risk management. Journal of Business & Economic Statistics 19(4), 465–474, 2001. URL https://doi.org/10.1198/07350010152596718.

FDIC Board of Governors of the Federal Reserve System, OCC. Revised guidance on model risk management. SR letter 26-2 / OCC bulletin 2026-13. Federal Reserve / OCC / FDIC supervisory guidance (April 17, 2026), 2026. URL https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm.

L. Breidenbach, C. Cachin, B. Chan, A. Coventry, S. Ellis, A. Juels, F. Koushanfar, A. Miller, B. Magauran, D. Moroz, S. Nazarov, A. Topliceanu, F. Tramèr, and F. Zhang. Chainlink 2.0: next steps in the evolution of decentralized oracle networks. Chainlink research whitepaper (April 2021), 2021. URL https://research.chain.link/whitepaper-v2.pdf.

P. F. Christoffersen. Evaluating interval forecasts. International Economic Review 39(4), 841–862, 1998. URL https://doi.org/10.2307/2527341.

CoinDesk. The market for tokenized equities has exploded by almost 3,000% in a single year. CoinDesk (Business), 2026. URL https://www.coindesk.com/business/2026/01/30/the-market-for-tokenized-equities-has-exploded-by-almost-3-000-in-a-single-year.

L. W. Cong, W. R. Landsman, D. Rabetti, C. Zhang, and W. Zhao. Tokenized stocks. Working paper, SSRN id 5937314 (December 2025), 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5937314.

S. Cuonzo and N. Deliu. Conformal prediction intervals with tail-specific guarantees. arXiv:2606.18199v1 [math.ST], 16 June 2026 (Sapienza University of Rome; MRC Biostatistics Unit, Cambridge), 2026. URL https://arxiv.org/abs/2606.18199.

P. Daian, S. Goldfeder, T. Kell, Y. Li, X. Zhao, I. Bentov, L. Breidenbach, and A. Juels. Flash boys 2.0: Frontrunning, transaction reordering, and consensus instability in decentralized exchanges. IEEE Symposium on Security and Privacy 2020; arXiv:1904.05234, 2020. URL https://arxiv.org/abs/1904.05234.

F. X. Diebold, T. A. Gunther, and A. S. Tay. Evaluating density forecasts with applications to financial risk management. International Economic Review 39(4), 863–883, 1998. URL https://doi.org/10.2307/2527342.

C. Duley, L. Gambacorta, R. Garratt, and P. Koo Wilkens. The oracle problem and the future of DeFi. BIS Bulletin No. 76, Bank for International Settlements (September 2023), 2023. URL https://www.bis.org/publ/bisbull76.pdf.

R. F. Engle and S. Manganelli. CAViaR: conditional autoregressive value at risk by regression quantiles. Journal of Business & Economic Statistics 22(4), 367–381, 2004. URL https://doi.org/10.1198/073500104000000370.

S. Eskandari, M. Salehi, W. C. Gu, and J. Clark. SoK: oracles from the ground truth to market manipulation. 3rd ACM Conference on Advances in Financial Technologies (AFT '21); arXiv:2106.00667, 2021. URL https://arxiv.org/abs/2106.00667.

Volcano Exchange. Post-mortem: tokenised-asset perpetual oracle propagation episode on hyperliquid. Volcano Exchange incident write-up (2025), 2025. URL https://volcano.exchange/blog/postmortem-tokenized-asset-perp-oracle-propagation.

Kamino Finance. Kamino is integrating xStocks powered by the chainlink data standard. Kamino governance forum (proposal thread), 2025a. URL https://gov.kamino.finance/t/kamino-is-integrating-xstocks-powered-by-the-chainlink-data-standard-to-enable-tokenized-equities-le 792.

Kamino Finance. –2026. scope oracle-type registry (‘oracle_type.rs‘). Kamino-Finance/scope GitHub repository (Rust source, on-chain program), 2025b. URL https://github.com/Kamino-Finance/scope/blob/master/programs/scope/src/states/oracle_type.rs.

Kamino Finance. Scope mainnet configuration — xStocks reserve mapping. Kamino-Finance/scope GitHub repository (mainnet config JSON, account ‘3NJYftD5sjVfxSnUdZ1wVML8f3aC6mp1CXCL6L7TnU8C‘), 2026. URL https://github.com/Kamino-Finance/scope/blob/master/configs/mainnet/3NJYftD5sjVfxSnUdZ1wVML8f3aC6mp1CXCL6L7TnU8C.json.

K. R. French. Stock returns and the weekend effect. Journal of Financial Economics 8(1), 55–69, 1980. URL https://doi.org/10.1016/0304-405X(80)90021-5.

I. Gibbs and E. J. Candès. Adaptive conformal inference under distribution shift. NeurIPS 2021; arXiv:2106.00170v3 [stat.ME], 2021. URL https://arxiv.org/abs/2106.00170.

I. Gibbs and E. J. Candès. Conformal inference for online prediction with arbitrary distribution shifts (DtACI). arXiv:2208.08401v3 [stat.ME] (Journal of Machine Learning Research), 2022. URL https://arxiv.org/abs/2208.08401.

I. Gibbs, J. J. Cherian, and E. J. Candès. Conformal prediction with conditional guarantees. Journal of the Royal Statistical Society Series B 87(4), 1100–1123; arXiv:2305.12616v4 [stat.ME], 2025. URL https://academic.oup.com/jrsssb/article/87/4/1100/8058684;https://arxiv.org/abs/2305.12616.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society, Series B 69(2), 243–268, 2007. URL https://doi.org/10.1111/j.1467-9868.2007.00587.x.

J. Gonzalo and C. W. J. Granger. Estimation of common long-memory components in cointegrated systems. Journal of Business & Economic Statistics 13(1), 27–35, 1995. URL https://doi.org/10.1080/07350015.1995.10524576.

J. Hasbrouck. One security, many markets: determining the contributions to price discovery. Journal of Finance 50(4), 1175–1199, 1995. URL https://doi.org/10.1111/j.1540-6261.1995.tb04054.x.

Hyperliquid HLP. JELLYJELLY perpetual incident report. Hyperliquid incident announcement (March 2025), 2025. URL https://hyperliquid.xyz/blog/jellyjelly-incident-mar-2025.

RWA.xyz / InvestAX. Q1 2026 Real-World asset tokenization market report. Industry market report (aggregator), 2026. URL https://investax.io/blog/q1-2026-real-world-asset-tokenization-market-report.

Kraken. xStocks remain largest provider of tokenized equities, surpass $25 billion in total transaction volume. Kraken Blog, 2026-02-19 (issuer/distributor primary source), 2026. URL https://blog.kraken.com/product/xstocks/25-billion-in-total-transaction-volume.

P. H. Kupiec. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives 3(2), 73–84, 1995. URL https://doi.org/10.3905/jod.1995.407942.

Chainlink Labs. –2026. chainlink data streams. Chainlink product documentation, 2024. URL https://docs.chain.link/data-streams.

Chainlink Labs. Data streams v10 ("tokenized asset") report schema. Chainlink product documentation; SDK source (authoritative pin), 2025a. URL https://docs.chain.link/data-streams/reference/report-schema-v10.

Chainlink Labs. Data streams v8 ("RWA standard") report schema. Chainlink product documentation; SDK source (authoritative pin), 2025b. URL https://docs.chain.link/data-streams/reference/report-schema-v8.

Chainlink Labs. Data streams v11 ("RWA advanced") report schema. Chainlink product documentation; SDK source (authoritative pin), 2026. URL https://docs.chain.link/data-streams/reference/report-schema-v11.

J. Luo, X. Xiong, Z. Li, H. Kang, X. Liu, W. J. Knottenbelt, and K. Tinn. SoK of RWA tokenization: A systematization of concepts, architectures, and legal interoperability. arXiv:2604.06608v2 [q-fin.GN], 14 April 2026 (McGill / Imperial College London / MBZUAI), 2026. URL https://arxiv.org/abs/2604.06608.

Nasdaq. Proposal for near-24-hour weekday trading of US-listed equities. Nasdaq announcement (Dec 2025), reported via CNBC, 2025. URL https://www.cnbc.com/2025/12/16/nasdaq-moves-to-near-24-hour-trading-some-say-thats-a-bad-idea.html.

Pyth Network. A proposal for price feed aggregation. Pyth Network blog / engineering post (documenting the deployed aggregation), 2021. URL https://www.pyth.network/blog/pyth-price-aggregation-proposal.

Pyth Network. Pyth primer: don't be pretty confident, be pyth confident. Pyth Network blog, 2022. URL https://www.pyth.network/blog/pyth-primer-dont-be-pretty-confident-be-pyth-confident.

Pyth Network. -09-25. blue ocean ATS joins pyth network: institutional overnight hours US equity data. Pyth Network blog (companion announcement on Blue Ocean's site), 2025. URL https://www.pyth.network/blog/blue-ocean-ats-joins-pyth-network-institutional-overnight-hours-us-equity-data;https://blueocean-tech.io/2025/09/25/pyth-blue-ocean-ats-joins-pyth-network-institutional-overnight-hours-u

Pyth Network. Pyth pro (formerly pyth lazer): enterprise-grade price data with customizable cadence. Pyth Network product documentation, 2026. URL https://docs.pyth.network/lazer(accessed2026-04-29; pagereframestheproductas"PythPro(formerlyPythLazer)").

Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. Supervisory guidance on model risk management. SR letter 11-7 / OCC bulletin 2011-12. Federal Reserve / OCC supervisory guidance, 2011. URL https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm.

Basel Committee on Banking Supervision. Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements. BCBS Publication 22 (January 1996), 1996. URL https://www.bis.org/publ/bcbs22.pdf.

S. Park, O. Bastani, and T. Kim. ACon²: adaptive conformal consensus for provable blockchain oracles. USENIX Security Symposium 2023; arXiv:2211.09330v3 [cs.CR], 2023. URL https://arxiv.org/abs/2211.09330.

UMA Project. UMA: an optimistic oracle with a data verification mechanism. UMA project documentation / blog post *Introducing UMA's Optimistic Oracle*, 2020. URL https://medium.com/uma-project/introducing-umas-optimistic-oracle-d92ce5d1a4bc.

SEDA Protocol. SEDA benchmark — session-aware oracle programs and the USA500 composite. SEDA Protocol product / documentation site, 2026. URL https://www.seda.xyz/;https://docs.seda.xyz/home/session-aware-oracle-programs;https://docs.seda.xyz/home/llms-full.txt.

A. Rafe and S. Das. Socio-conformal calibration in complex survey data: marginal validity is not enough for subgroup reliability. arXiv:2605.05562v1 [stat.ME], 7 May 2026 (Texas State University), 2026. URL https://arxiv.org/abs/2605.05562.

RedStone. RedStone live: real-time market data for 24/7 markets. RedStone blog / product documentation (March 2026), 2026. URL https://blog.redstone.finance/2026/03/30/redstone-live-real-time-data-built-for-the-markets-that-never-sleep/.

Y. Romano, E. Patterson, and E. J. Candès. Conformalized quantile regression. Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html;arXiv:1905.03222.

M. Schmitt. Taming tail risk: regime-weighted conformal calibration for nonstationary Value-at-Risk. arXiv:2602.03903v2 [q-fin.RM], submitted 3 February 2026, revised 13 July 2026 (University of Oxford), 2026. URL https://arxiv.org/abs/2602.03903.

U.S. Securities and Exchange Commission. Orders granting accelerated approval for extended trading hours (24X national exchange; NYSE arca). SEC release / SIFMA extended-trading-hours tracker, 2025. URL https://www.sifma.org/issues/market-structure/extended-trading-hours.

H. R. Stoll and R. E. Whaley. The dynamics of stock index and stock index futures returns. Journal of Financial and Quantitative Analysis 25(4), 441–468, 1990. URL https://doi.org/10.2307/2331010.

Switchboard. Switchboard v3 documentation. Switchboard product documentation / Breakpoint 2023 presentation, 2023. URL https://docs.switchboard.xyz/docs/switchboard/readme/designing-feeds/oracle-aggregator.

Tellor. Tellor whitepaper. Tellor project whitepaper (April 2023 revision), 2023. URL https://tellor.io/wp-content/uploads/2023/04/Tellor-Whitepaper_4_2023.pdf.

V. Vovk, D. Lindsay, I. Nouretdinov, and A. Gammerman. Mondrian confidence machine. On-Line Compression Modelling project working paper, 2003. URL http://www.alrw.net/old/04.pdf(canonicalintroduction; alsodiscussedinvovk-book-2005Âğ4.5).

V. Vovk, A. Gammerman, and G. Shafer. Algorithmic learning in a random world. Springer, Boston, MA; ISBN 978-0-387-00152-4, 2005. URL https://doi.org/10.1007/b106715.

D.-Y. Wang, G. Liang, K. Zhang, and Q. Zhu. Reliable real-time Value-at-Risk estimation via quantile regression forest with conformal calibration. arXiv:2602.01912v1 [stat.ML], 2 February 2026 (Renmin University of China; City University of Hong Kong), 2026. URL https://arxiv.org/abs/2602.01912.

C. Xu and Y. Xie. Conformal prediction interval for dynamic time-series (EnbPI). Proceedings of the 38th International Conference on Machine Learning (ICML 2021), PMLR 139, 2021. URL https://proceedings.mlr.press/v139/xu21h.html.

M. Zaffran, A. Dieuleveut, O. Féron, Y. Goude, and J. Josse. Adaptive conformal predictions for time series. ICML 2022; arXiv:2202.07282v1 [stat.ML], 2022. URL https://arxiv.org/abs/2202.07282.

T. Zhong. Proxy-reliance control in conformal recalibration of one-sided Value-at-Risk. arXiv:2603.22569v1 [q-fin.RM], 23 March 2026 (University of Southern California), 2026. URL https://arxiv.org/abs/2603.22569.