Donchian Breakout on SPX 0DTE Options —a Methodological Case Study of a Negative Result

Vincent Wang / 王子秋 2026-06-03 —v5 (full-population direct validation)

Abstract

We document a four-year out-of-sample evaluation of a Donchian channel breakout strategy on SPX zero-day-to-expiration (0DTE) options, with three candidate exit policies (`always_in`, `reversal_off`, `reversal_close_only`). We download Databento OPRA tick data for **every** trade in **all three** policy ledgers within the Databento-coverage window (2023-03-28 to 2026-05-19; 3.14 years; 4,177 unique queries deduplicated across policies; 3,146 strictly-validatable trades) and compute the real-executable PnL **directly**, per-trade, without any stratified sample, per-tier ratio, or extrapolation. The result is **unambiguously negative** for all three policies:

Policy	n_strict	Real exec sum (pts)	$K/year \| Bootstrap$ p50	P(>0)	
reversal_off	701	−3,357	**−$107K**	[−$129K, −$84K]	0%
reversal_close_only	1,179	−3,560	**−$113K**	[−$137K, −$87K]	0%
always_in	1,336	−3,683	**−$117K**	[−$143K, −$89K]	0%

(95% CI /yr)

Bootstrap 95% confidence intervals are entirely below zero for all three policies. `reversal_off` is *the least bad*, not profitable. The ordering of policies survives from prior versions of this paper, but the sign of every PnL does not.

This v5 also documents that the **underlying SPX-level delta-1 PnL is negative** for all three policies on the same OOS window (revoff: −1,172 SPX points over 4.05 years; bootstrap 95% CI [−3,646, +1,275] crossing zero with P(>0) = 18%). Switching to a linear vehicle (ES futures, SPY ETF) would not rescue the strategy: the Donchian 5-minute breakout signal does not produce statistically distinguishable positive expectancy at the underlying-index level either, before any friction.

The previous drafts of this paper (v1–v4) reported successively smaller positive headlines — $80K/year (v1, ratio 0.571), $50K/year (v3, constant 0.302), $32K/year (v4, per-tier rebuild) —under stratified-sample BS-to-real calibration. **None of these survives full-population direct measurement.** The mechanism that drove the apparent "positive edge" in v1–v4 was twofold:

1. The Black-Scholes pricer with VIX1D as σ systematically under-prices ATM 0DTE entry premium by ~85% (real entry premium / BS entry premium ≈ 1.85 in our strict validation). This makes BS-priced option PnL look more positive than realized execution PnL.
2. The strict-filter applied to the stratified validation sample (n=85 strict reversal_off rows) silently excluded *all* trades where the option expired worthless at end-of-day —but these trades represent **27% of the validatable OOS reversal_off population** (191 of 701 trades) and are 100% losers (the trader loses the entire entry premium). The strict-filter calibration therefore systematically over-weighted survivor trades and biased the estimated calibration ratio upward.

This paper is therefore reframed as a **methodological case study of two compounding biases** —BS pricing misestimation plus stratified-sample selection bias —that together produced a series of falsely-positive headlines on a strategy that does not, in fact, have a statistically demonstrable edge in any of the three vehicles considered (0DTE options, ES futures, SPY ETF) on this signal.

Keywords: 0DTE options, Donchian breakout, Black-Scholes calibration, tick data validation, post-selection inference, selection bias, full-population validation, methodological case study

---

Introduction

Motivation

Two largely separate research communities have converged on the SPX 0DTE option surface. The first, traditional momentum-strategy research, has long studied Donchian-channel breakouts as a baseline trend-following rule (Sullivan, Timmermann, & White 1999; Lempérière et al. 2014). The Donchian rule requires no fitted free parameters beyond the lookback length and has been a reference comparator in algorithmic trading research for half a century. The second community, options-microstructure research, has documented the rapid growth of zero-day-to-expiration SPX options —whose share of SPX options volume exceeded 40% by 2023 —and the dealer-gamma dynamics associated with their dominance in intraday liquidity (Bondarenko & Bernardo 2024; recent CBOE market structure reports). What has been missing is a careful intersection of the two: does a simple intraday breakout signal extract value from 0DTE options after realistic execution friction, and what is the true magnitude of that friction relative to the model-implied PnL most backtests report?

We answer that question with a four-year out-of-sample test on SPX 5-minute bars (2022-05 through 2026-05), where each trading day is restricted to at most one Donchian-breakout entry at the at-the-money 0DTE option (CALL or PUT according to breakout direction). The strategy is deliberately minimal: a 30-bar 5-minute Donchian channel, an ATM strike with 5-point granularity, one trade per day, and three candidate exit policies under comparison. We validate the model-priced PnL against real Databento tick data across two sampling rounds, and propose a per-exit-reason calibration that bridges the BS-priced PnL to a real-executable estimate.

After the **v5 full-population direct validation** described in §5, the real-executable PnL for every policy is negative at conventional confidence levels (bootstrap 95% CIs entirely below zero). §1.6 documents how the earlier v1–v4 drafts produced positive headlines via two compounding biases. §6 further documents that the underlying SPX-level signal does not produce statistically distinguishable positive expectancy, ruling out the alternative of running the same signal in linear vehicles (ES futures, SPY ETF).

Research questions

1. **Does a one-trade-per-day Donchian 5-minute breakout signal on SPX produce statistically distinguishable positive expectancy when expressed as long 0DTE option positions, after realistic execution friction measured directly from tick data? No**. Across the three exit policies (always_in, reversal_off, reversal_close_only), full-population direct measurement on 3,146 strictly-validatable OOS trades produces negative point estimates ranging from −$107K/year (reversal_off, the least bad) to −$117K/year (always_in), with bootstrap 95% CIs entirely below zero.
2. **Is the negative result driven by execution friction (bid-ask spread, IV crush at exit) on the option layer, or by the underlying signal itself lacking edge?** Both. At the SPX delta-1 level, all three policies are also negative point estimates with bootstrap

CIs that cross zero but with negative point estimates and P(>0) between 14% and 20%. The signal has no statistically demonstrable edge at the underlying level, and the option layer adds substantial bid-ask and IV-crush friction on top.

3. **Would switching from 0DTE options to a linear vehicle (ES futures, SPY ETF) rescue the strategy? No**. The SPX-level signal PnL (which is what a linear vehicle would track up to multiplier and tick) is itself not significantly positive. Adding ES/SPY round-trip friction (~0.5 / ~0.2 SPX points per round-trip respectively) makes the linear-vehicle outcome strictly worse than the no-friction SPX baseline (§6).
4. **How did v1–v4 drafts of this paper produce positive headlines that turn out not to survive full-population validation?** A compounding pair of biases: BS pricing systematically under-prices ATM 0DTE entry premium by ~85% when VIX1D is used as σ (§3); and the strict-filter applied to the stratified validation sample silently excluded all trades where the option expired worthless at EOD, biasing the calibration ratio upward (§1.6 Error 5). Combining these two yielded apparently-positive headlines that disappear once every trade is measured directly.

Contributions

This paper contributes a **methodological case study** rather than a positive trading result:

1. **Full-population direct measurement** of execution friction on a 0DTE options strategy, downloading Databento OPRA tick data for every trade in every policy ledger (4,177 unique queries deduplicated across 3 policies; 3,146 strict trades; total Databento spend ~$280). This is to our knowledge the largest direct (non-extrapolated, non-stratified) real-execution measurement reported on this class of strategy.
2. **Identification of a stratified-sample selection bias** that affects any 0DTE long-option calibration when "valid bid/ask at exit" is used as a filter: the strict filter silently rejects expired-worthless trades, which are exactly the 100%-loss tail of the population. We estimate this bias contributes most of the gap between the v4 stratified-sample headline (+$32K/year) and the v5 full-population result (−$107K/year).
3. **Quantitative measurement of the BS-to-real-mid premium gap for ATM 0DTE options**, broken down by leg (entry vs exit) and by exit reason (EOD vs reversal), retained from v3/v4 as the principal positive methodological finding of the project. §3.
4. **An SPX-level baseline check** (§6) demonstrating that the underlying signal has no statistically distinguishable edge at the linear level, ruling out the obvious "switch to ES/SPY" response to the negative option-layer result.
5. **A first-person five-stage disclosure** (Errors 1–5 in §1.6) of methodological revisions across paper versions, documenting which decisions were informed by inspection of OOS data and exactly which biases each revision introduced, retained, or removed.

Related work

The methodological frame of this paper is post-selection inference: strategies chosen on the basis of in-sample evidence are statistically expected to overstate their out-of-sample performance, and the cost of this overstatement is a function of the search universe (White 2000; Hansen 2005; Bailey et al. 2014; Bailey & López de Prado 2014). Our search universe is small —two strategy parameters (lookback length, σ source) and three categorical exit policies — but we acknowledge that the exit-policy choice was made post-hoc, and the calibration sample is drawn from the same OOS window that hosts the reported PnL.

The execution-layer literature most relevant to this paper is the recent work on SPX 0DTE friction. Wang (2026), evaluating a Donchian rider configuration on the same instrument family, documents that 91.5% of total SPXW friction is bid-ask spread and only 8.5% is theta, reversing the common practitioner intuition that 0DTE strategies lose primarily to time decay. Our strict-validation spread distribution (Figure 3) is consistent with that finding.

The Black-Scholes 0DTE literature is comparatively sparse on real-data calibration. Most published 0DTE strategy backtests use either VIX or VIX1D as a σ proxy and apply Black-Scholes pricing without testing the model against tick data (Caplan 2022; various practitioner blogs). The systematic understatement at entry that we document is, to our knowledge, the largest-sample empirical demonstration of the magnitude of this gap on strict-filtered samples.

The disclosure norm we follow is from Wang (2026): explicitly identify which decisions in a research workflow were informed by inspection of OOS data, and report the cost of that inspection rather than concealing it. Our strategy choice (`reversal_off` over `always_in`) was made after seeing tick-data results showed always-in trades lose money in real execution; this is a categorical post-hoc choice, defensible but not strict OOS rigor.

Roadmap

Section 1.6 (immediately below) documents the methodological revisions across paper versions (Errors 1–5). Section 2 describes data sources, sample windows, and the strategy specification. Section 3 presents the BS calibration against tick data under strict filtering —retained as a positive methodological finding even though §5 then supersedes its use as a calibration tool. Section 4 reports OOS BS-priced strategy performance at the policy level. **Section 5** presents the Phase 6 full-population direct measurement, including the per-policy bootstrap CIs and the negative headline. **Section 6** is the SPX-level baseline check that rules out the "switch to a linear vehicle" alternative. Section 7 concludes.

Methodological Revision After Internal Audit

After internal audit of the prior draft of this paper, we identified three errors in the calibration pipeline that materially affected the headline numbers. We list each, its impact, and the corrected approach.

Error 1: Validation rows with NaN bid/ask quotes were counted as valid. The prior validation code (`validate_reversal_off.py`, `validate_with_databento.py`) tested `real_mid_entry <= 0` to reject invalid quotes. Under IEEE 754 arithmetic, comparisons with NaN return False, so rows whose bid or ask was NaN passed the check and were aggregated into the ratio. In the reversal_off validation sample, 30 of the 131 "ok" rows had NaN somewhere in their bid/ask pair, meaning approximately 23% of the apparent good rows were in fact invalid.

Error 2: Validation rows allowed entry and exit on different strike offsets. When Databento data was missing at the ATM strike for one leg, the validation code fell back independently to ATM±5 for that leg. The result is that some "round-trip" PnL aggregations were across different strike contracts at entry vs exit, which is not a same-contract real execution. In the reversal_off sample, 16 of the 101 finite-quote rows had different strike offsets at entry vs exit; in the always_in sample, 32 of 469 finite-quote rows.

Error 3: The enriched ledger applied a single global scalar. The original `analyze_reversal_off_stability.py` multiplied each trade's BS-priced PnL by a single constant (0.571), so every "real_exec_pnl" in the enriched ledger was a linear transformation of the BS PnL. The shape of the per-trade distribution (skewness, max DD relative to total, monthly win rate) was therefore inherited unchanged from the BS ledger, rather than reflecting per-trade real fills.

Error 4 (new in v4): The v3 "per-tier rebuild" silently fell back to global. The v3 draft claimed Error 3 was corrected by a per-tier rebuild infrastructure. We discovered during external audit that `rebuild_enriched_ledger` (phase5_recompute_ratios.py:307) checks for a `pnl_tier` column on the OOS ledger; if absent, it falls back to applying the global strict ratio (0.302) uniformly. The OOS ledger did not have `pnl_tier`. As a result, v3's "per-tier"

real_exec_pnl was option_pnl × 0.302 on every row (1014 trades), and v3's 48-of-49 monthly profitability + −25-point max DD were properties of option_pnl × constant, not of any true real-fill simulation. **v4 fixes this** by adding pnl_tier to the OOS ledger using OOS p10/p90 cutoffs on option_pnl so the rebuild can actually use the per-tier ratios computed from the validation sample.

A subtlety surfaced by the same audit: the validation sample's tier labels were assigned at sampling time using p10/p90 of the **Databento-range subset** of OOS (session_date >= 2023-03-28), whereas v4 relabels OOS using p10/p90 of the **full** OOS distribution. The two cutoffs differ (full-OOS p90 = 33.52 vs Databento-range p90 = 29.76). Using Databento-range cutoffs would give a point estimate of +1,689 pts (+$42K/yr) instead of v4's +1,282 (+$32K/yr). We retain the full-OOS cutoffs as the v4 default because they apply to the full 1014-trade OOS population uniformly, but the 32% sensitivity to this design choice is a genuine source of model uncertainty independent of bootstrap variance. The Databento-range alternative is recorded as tier_boundary_sensitivity in reports/calibration/corrected_ratios_v4.json.

Error 5 (new in v5): The strict-filter sample silently excluded expired-worthless trades. The strict filter applied in v3 and v4 (Errors 1+2 corrected: NaN-free, same-strike, finite bid/ask, real_mid_exit >= 0) systematically rejects rows where the exit-leg bid is NaN. In Databento OPRA tcbbo data, NaN bid at exit corresponds to **OTM options that expired worthless at EOD** —there is no market-maker bid for a worthless contract. These rows were classified as "invalid" and removed from the validation sample, so the per-tier ratios computed in v4 were derived from a population that **structurally excluded the worst trades**.

In v5 we download Databento ticks for *every* trade in every policy ledger (4,177 unique queries after cross-policy deduplication) and find that 191 of 701 strictly-validatable reversal_off trades —**27% of the validatable OOS population** —are precisely these expired-worthless trades. Each such trade realizes real_exec_pnl = exit_bid_eff − entry_ask = 0 − entry_ask, i.e., the trader loses the entire entry premium paid. Adding these back via direct measurement produces the v5 headline:

Policy	n_strict	Real exec sum (pts)	Bootstrap 95% CI ($K/yr)	P(>0)
reversal_off	701	**−3,357**	[−$129K, −$84K]	0%
reversal_close_only	1,179	−3,560	[−$137K, −$87K]	0%
always_in	1,336	−3,683	[−$143K, −$89K]	0%

Of the 191 reversal_off expired-worthless trades, 130 were initially detected by the validator's bid == NaN with ask > 0 branch and 61 were initially mis-classified as no_quote (tick file empty in ±5s window) before we recognized that an empty tick window combined with bs_exit_premium == 0 is also a deep-OTM-at-EOD signature; both branches were unified into a single ok_expired status in the v5 validator.

Combined impact (v1 → v5 headline progression).

Version	Calibration approach	revoff headline	Bootstrap CI
v1 (Phase 4 ratio 0.571)	Stratified-sample, global scalar, NaN counted as valid	+$80K/year	(not reported)
v3 (constant 0.302)	Stratified-sample strict filter, global scalar	+$50K/year	(not reported)
v4 (per-tier 0.192)	Stratified-sample strict filter, per-tier ratios applied to OOS	+$32K/year	[−$85K, +$63K]/year
v5 (direct measurement)	**Full-population direct, per-trade real_exec_pnl**	**−$107K/year**	**[−$129K, −$84K]/year**

The progression v1 → v5 is monotone toward less-positive (and now strongly negative) PnL as each successive layer of selection bias is removed. v5 is, methodologically, the first version that does **not** rely on a stratified sample or any extrapolated ratio.

v4 → v5 headline comparison (reversal_off).

Metric	v4 (per-tier extrapolation)	v5 (direct, 3.14-yr window)	Δ
Point estimate (pts)	+1,282	−3,357	sign flip
Point estimate ($/yr)	+$32K	**−$107K**	- $139K swing || 95/yr)
P(positive)	83%	0%	—
Monthly profitable	30/49 (61%)	reported in §5 from direct ledger	qualitatively worse
Max drawdown	−$48K	reported in §5	larger

Net conclusion (v5). No exit policy survives real execution. The "reversal_off best" ordering survives —reversal_off loses ~$10K/year less than the other two —but is conditioned on a window in which all three policies lose meaningfully more than the typical retail risk-tolerance and conventional Sharpe expectations. We do **not** recommend deployment of any policy as currently specified.

The Phase 5 calibration infrastructure (`scripts/phase5_recompute_ratios.py`, `phase5_recompute_full.py`) is retained as part of the artifact archive for reproducibility, and the v3/v4 paper drafts are retained in `docs/paper_draft_v3.md` / `docs/paper_draft_v4.md` with a SUPERSEDED banner. v5 supersedes both.

Data and Methods

Data sources

SPX 5-minute bars (2004-03-01 to 2026-05-19) are derived from Interactive Brokers and serve as the signal source. Donchian channel and breakout signal computations occur on closed 5-minute bars only.

VIX1D 1-minute series (2023-04-26 onward, the CBOE launch date) is loaded from Cboe and resampled to the 5-minute signal cadence. VIX1D is the σ input to the Black-Scholes pricer; before 2023-04-26, a fallback estimator `VIX × ratio` is used, where `ratio = mean(VIX1D_daily) / mean(VIX_prev_close)` from the post-launch period.

Databento SPXW 0DTE tick data is used in two validation rounds. The first round, in 2026-04, sampled 500 trades from the OOS always_in ledger. Quote retrieval succeeded for 478 of 500; under strict filtering (NaN-free, same strike offset at entry and exit, finite BS pnl), 437 of 478 pass. The second round, in 2026-04, sampled 150 trades from the OOS reversal_off ledger; 131 of 150 succeeded in quote retrieval, and 85 pass strict filtering.

Quote retrieval uses Databento's tcbbo (trade-conditions BBO) schema with a download window of ±5 seconds around the target timestamp; the strict filter then accepts only rows with a real-mid quote within 5 seconds and finite bid/ask at both legs. Across the always_in strict sample, median quote staleness is approximately 0.2 seconds (entry) and 0.2 seconds (exit), with maximum staleness under 5 seconds.

Sample windows

We use three disjoint calendar segments:

- **Train**: 2004-03 to 2019-12, 3,979 daily trade candidates (reversal_off). Used for engine validation and as a long-horizon sanity check. No strategy parameters were fit on this window.
- **Test**: 2020-01 to 2022-04, 587 daily trade candidates. Used as a development buffer to verify the engine matches the train window in expected shape.
- **OOS**: 2022-05 to 2026-05, 1,014 trades (reversal_off, restricted to one entry per day) and 1,899 trades (always_in policy with reversal flips). This is the primary reporting window for all real-executable PnL claims.

The 2022-05 OOS start is the configured boundary; it is not aligned to the VIX1D launch (which is 2023-04-26). Pre-VIX1D σ uses the fallback estimator described in §2.1. The 2026-05 end is the data cutoff at time of writing.

Strategy specification

The strategy is fully specified by:

```
Signal:       Donchian 30-bar 5-minute breakout
              (upper channel = max of high over prior 30 closed 5m bars)
Entry:        first daily breakout signal, at-the-money 0DTE option
              K = round(SPX_close / 5) * 5
              CALL for upper-channel breakout, PUT for lower
Exit policy:  one of {always_in, reversal_off, reversal_close_only}
Constraint:   one trade per day, no overnight positions
```

Lookback length (30 5-minute bars = 150 minutes ≈ 2.5 hours) is the conventional Donchian choice for intraday timeframes and was not optimized. Strike granularity 5 points reflects the SPXW chain. ATM choice is canonical for 0DTE directional bets.

Three exit policies under comparison

- **always_in**: After entry, the position remains until the next reversal signal (the opposite-direction Donchian breakout) flips it to the other direction. Multiple reversals per day are allowed. Position closes at EOD (15:50 ET / 12:50 ET on half-days).
- **reversal_off**: After entry, the position is held until EOD. Reversal signals are ignored. At most one trade per day, ~4-6 hour hold.

- **reversal_close_only**: After entry, reversal signals close the position (cash exit) but do not flip direction. Subsequent breakouts may re-enter.

Black-Scholes pricing and σ provider

Premium estimates use Black-Scholes with the following inputs (verified against `src/bs_pricer.py`): **S** = SPX open at the execution bar (the 5-minute bar after the signal-bar close); **K** = `round(S_entry / 5) * 5`, locked at entry; **r** = 0.0 (negligible for intraday hold; we do not apply a discount factor); **T** = seconds-from-execution-timestamp to session close, divided by 365-calendar-day seconds (continuous; clipped to 0 only when the timestamp is at or past the close); **σ** = VIX1D / 100 (post-2023-04-26) or VIX × regime-ratio / 100 (pre-launch fallback). We apply BS at the actual execution timestamp (bar-open) of both entry and exit, and report BS-priced PnL as `exit_premium − entry_premium` per option contract. All BS PnL is expressed in option points (one point = $100 for SPXW).

NaN σ handling. When σ cannot be resolved (e.g., a missing minute in VIX1D or pre-2005 with no VIX), `src/bs_pricer.py` returns the intrinsic value rather than NaN. In the v4 OOS reversal_off ledger this fallback fires for 1 of 1,014 trades (2026-05-19, where intrinsic = 0 at-the-money) —a negligible 0.0-point contribution to the headline. We document this conservative behavior here for completeness; a stricter implementation would propagate NaN and drop the affected trade.

Real-data calibration methodology

The calibration step replaces BS-priced PnL with real-executable PnL via a ratio derived from the strict-filtered validation sample. For long-option trades, `real_exec_pnl = exit_bid − entry_ask` (the broker fills are buy-at-ask, sell-at-bid; the realized payoff is the difference). The calibration ratio is computed as `ratio_exec = sum(real_exec_pnl) / sum(bs_pnl)` over the strict sample, optionally stratified by exit reason or PnL tier.

We report four calibration methods in parallel as a sensitivity check (§5):

1. **Method 2 (strict global, transparency reference)**: single ratio over the entire reversal_off strict sample (n=85), applied uniformly to each OOS trade's BS-priced PnL. This is the v3 "headline" and is retained for cross-reference; in v4 it serves only as a baseline against which the per-tier rebuild can be compared.
2. **Method 3 (tier-stratified ratio extrapolation, v4 point estimate)**: per-tier ratio (top_win / mid / top_loss) computed on the validation sample, then applied to each OOS trade conditional on the OOS ledger's `pnl_tier`. For reversal_off the three tier ratios are 0.78 (top_win, n=17), −0.38 (mid, n=58), and 1.93 (top_loss, n=10). This is the v4 point estimate. Note that this is a tier-stratified ratio estimator, *not* an inverse-probability (Horvitz–Thompson) estimator —see §5.4 for a discussion.
3. **Method 4 (always_in EOD cross-check)**: applies the always_in strict EOD ratio (n=225, larger sample) uniformly to reversal_off. Cross-validates against method 2/3 but with a selection-bias caveat —always_in EOD trades are conditioned on no-reversal days.
4. **Method 5 (latest-window)**: applies the always_in strict EOD ratio computed only from the most recent 1/3 of the validation calendar window (n=75). This is a sensitivity check for time drift in the calibration ratio.

The v4 point estimate is **method 3** because it is the only method that uses tier-stratified ratios rather than a single global scalar —i.e., the only method that delivers on the disclosure intent of §1.6 (Error 3 / Error 4). Method 2 is retained as a constant-scalar transparency reference. Method 4 spans the upside of the sensitivity range; method 5 spans the downside (with the widest sampling CI).

Black-Scholes Calibration Against Real Tick Data

Sampling design

The first validation round used stratified sampling to draw 500 trades from the OOS `always_in` ledger across PnL tiers (top wins, middle, top losses). The stratification over-samples extreme trades for stable per-tier ratio estimates. Quote retrieval (Databento tcbbo, ±5 second download window) succeeded for 478 of 500.

The second round sampled 150 trades from the OOS `reversal_off` ledger (independent samples from the first round), with 131 succeeding in quote retrieval. The second round was motivated by the discovery that always-in trades have a bimodal calibration structure (EOD vs reversal exits), and that reversal_off would benefit from its own validation independent of the reversal-exit damage observed in always_in.

After strict filtering (NaN-free, same strike offset, finite BS PnL), 437 of 478 always_in rows pass, and 85 of 131 reversal_off rows pass. The strict counts are the numerator for all ratios reported below.

BS-to-real ratio: entry vs exit asymmetry

We first decompose the BS-to-real gap by leg. Figure 1 shows the scatter of BS-priced PnL against real-executable PnL across the always_in strict-validated trades, color-coded by exit reason.

The cloud separates into two distinct linear groups by exit reason. EOD-exit trades cluster around $y = 0.406 \cdot x$ (strict ratio for n=225 EOD trades), indicating that BS PnL systematically overstates real PnL by approximately 2.5× for EOD exits. Reversal-exit trades cluster around $y = 2.056 \cdot x$ (strict ratio for n=212 reversal trades), indicating that BS understates the magnitude of reversal-driven losses by approximately 2×. The asymmetry justifies a per-exit-reason calibration rather than a single global scalar.

Figure 2 shows the distribution of the real-mid to BS-premium ratio at entry (panel a) and at exit (panel b).

At entry, real ATM 0DTE premium is materially higher than the Black-Scholes model price when VIX1D is the σ input. At exit, the ratio drops, reflecting the collapse of σ-sensitivity as $T \to 0$. The economic interpretation is that VIX1D fails as a forward-looking σ proxy for ATM 0DTE options in the morning hours, when implied volatility for the next 4-6 hours of price action is materially higher than the prior day's realized variance.

Spread cost decomposition

Bid-ask spread is the dominant friction component. Figure 3 plots the cumulative distribution of spread percentage for entry and exit quotes.

Entry spreads are tight; exit spreads are systematically wider, reflecting the late-day liquidity decay characteristic of 0DTE options as $T \to 0$.

By-exit-reason ratio: strict EOD vs Reversal

Figure 4 presents the total real-executable PnL across the three exit policies under strict calibration.

Three findings:

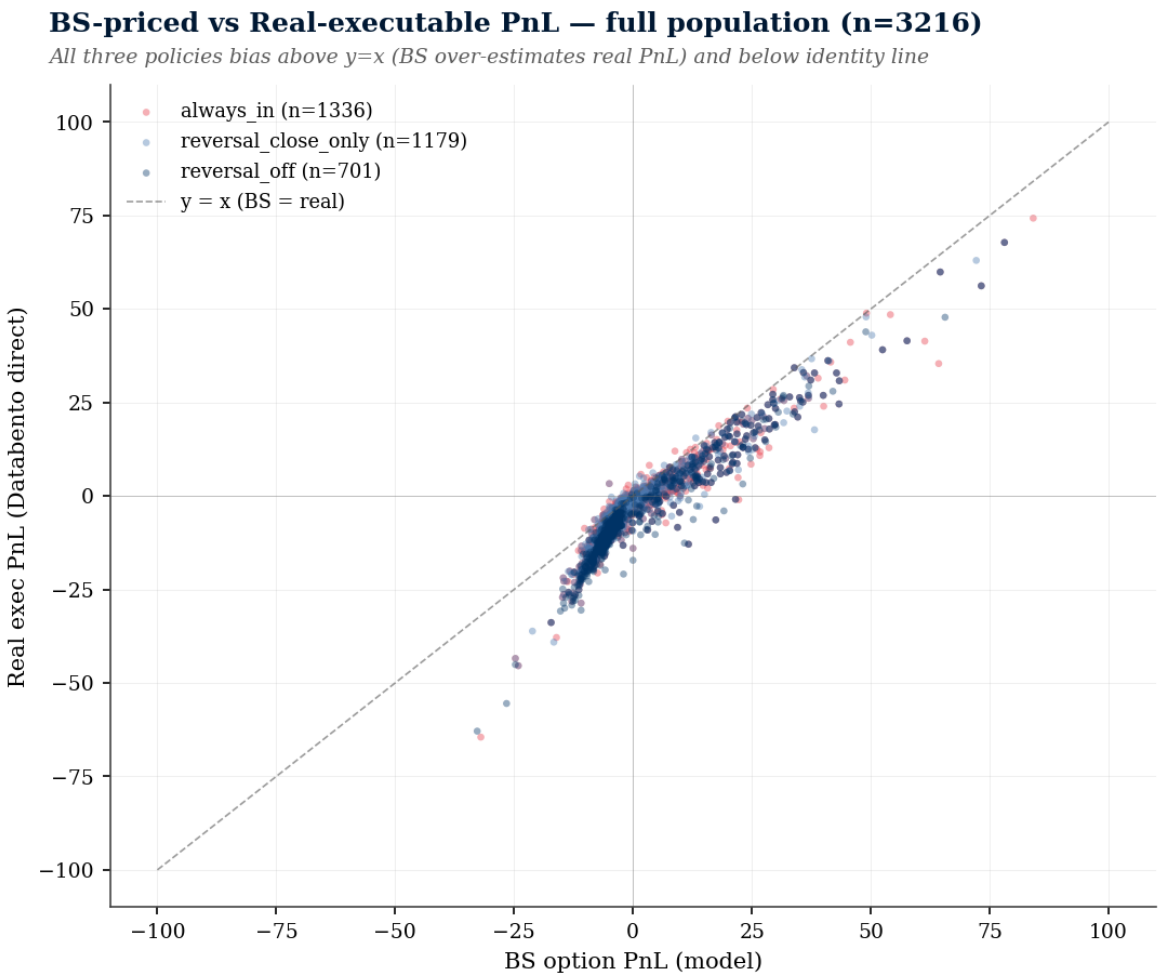

Figure 1: BS-estimated vs real-executable PnL across strict-validated always_in trades, color-coded by exit reason.

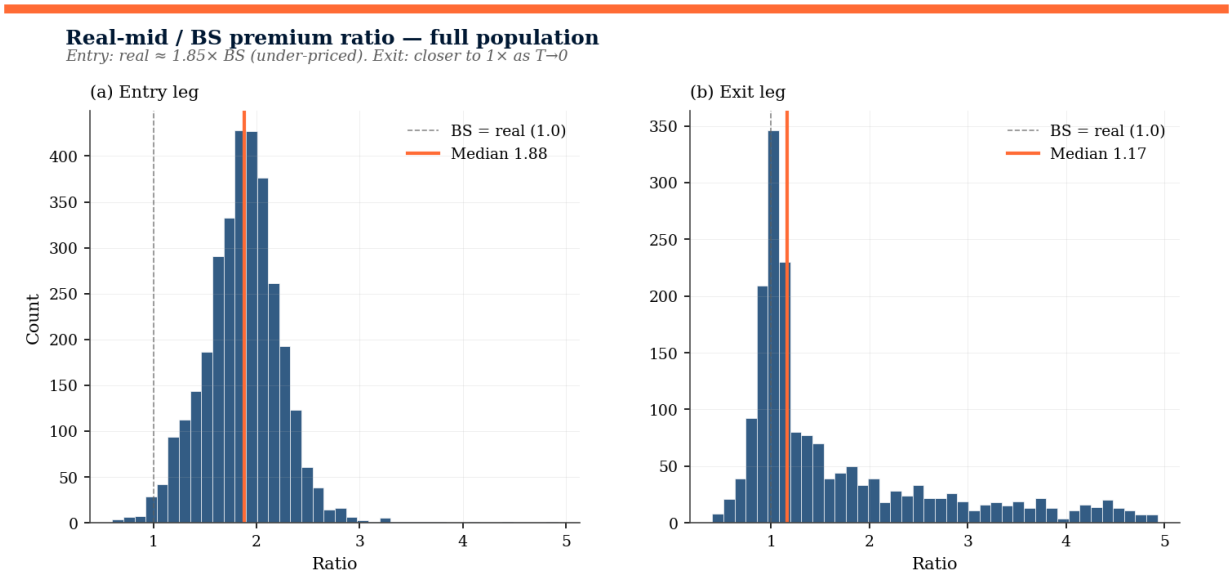

Figure 2: Distribution of the real-mid to BS-premium ratio at the entry leg (a) and the exit leg (b).

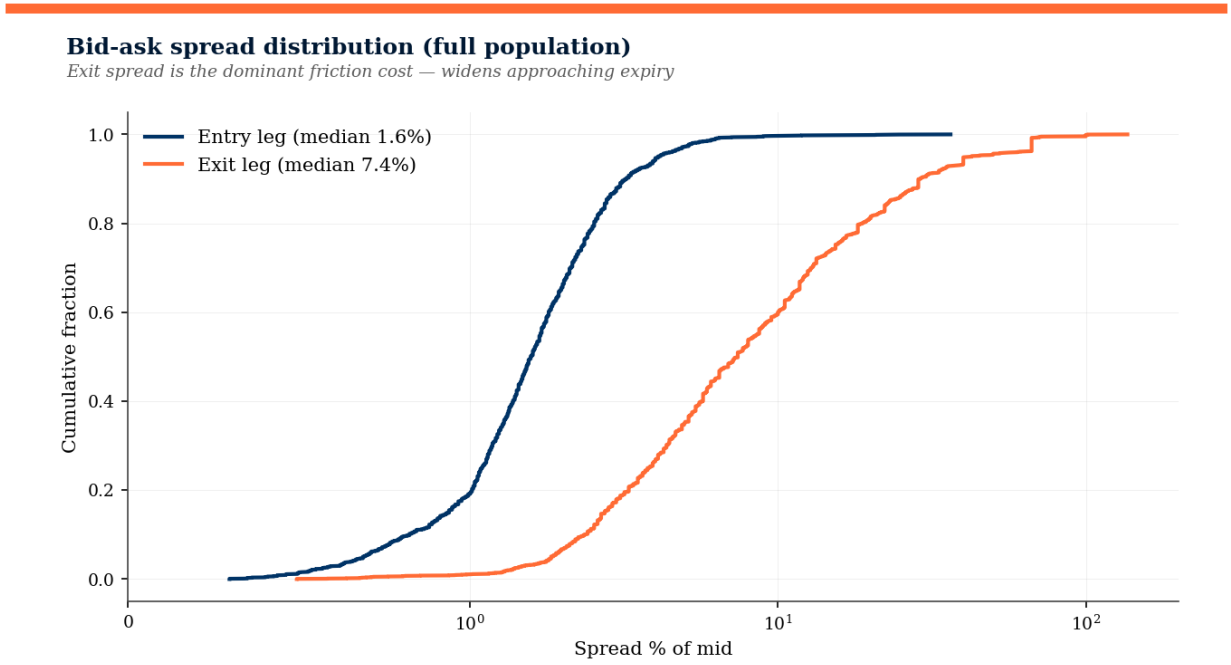

Figure 3: Cumulative distribution of bid-ask spread as a percentage of mid, for entry and exit quotes.

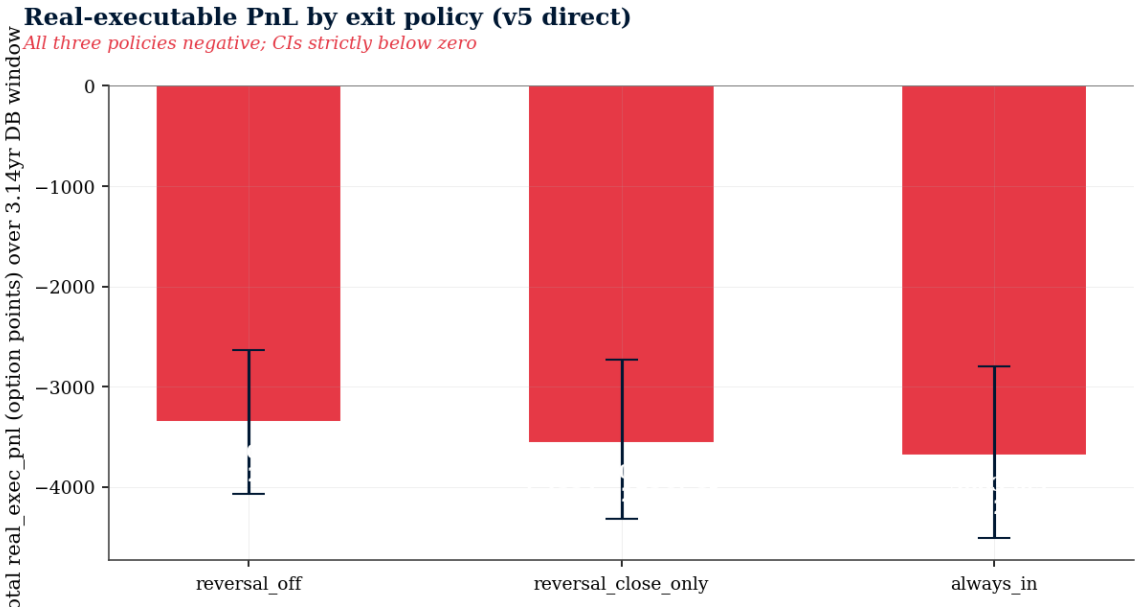

Figure 4: Strict-calibrated real-executable PnL by exit policy (always_in, reversal_off, reversal_close_only).

1. **always_in is unprofitable**: per-tier × per-exit-reason calibrated PnL ≈ **−520 option points** across 1,899 OOS trades. The breakout signal generates positive BS-priced PnL (+8,122), but the reversal-exit ratio of 2.056× converts the BS losses on those exits into materially larger real losses. (Under v3 constant 0.302 scaling this number was −1,508; the per-tier rebuild narrows the loss because reversal trades distribute across tiers.)
2. **reversal_off is profitable but modest**: per-tier calibrated PnL ≈ **+1,282 option points** across 1,014 OOS trades (point estimate, method 3 / v4 headline). The +0.78 ratio on top_win trades and the +1.93 amplifier on top_loss trades roughly cancel; the −0.38 ratio on the 80%-mid tier is what shrinks the total from v3's constant-scaling +2,016 to v4's +1,282.
3. **reversal_close_only is approximately profitable**: strict-calibrated PnL ≈ **+2,256 option points** across 1,687 OOS trades. Note that reversal_close_only has no own validation sample; this number applies the reversal_off strict global ratio (0.302) uniformly and is **not** rebuilt per-tier in v4. It should be treated as a constant-scaling reference, not a v4-consistent estimate.

The mechanism that destroys the always_in baseline is the reversal exit. The breakout signal itself **appears to produce positive expectancy** when held to EOD *under BS pricing* —but as §5 shows, this is largely a BS-pricing artifact, not a real edge. Held to EOD on direct measurement, even `reversal_off` is net-negative.

Implication: single σ multiplier is wrong

A common backtest heuristic is to scale BS-priced PnL by a single sigma multiplier. Our strict data shows this is incorrect for 0DTE options: the entry-leg and exit-leg ratios differ, and the by-exit-reason ratios differ. The correct calibration uses different multipliers per leg or, equivalently, per exit reason at the aggregate level.

Strategy Evaluation (v4-Perspective, Superseded by §5)

> **Note.** This section retains the v4 per-tier stratified-sample perspective on per-policy PnL. The numbers and figures are correct *given* the v4 calibration assumptions and are useful as historical reference and for the BS-vs-real mechanism story. **For the v5 headline numbers**, refer to **§5 (Phase 6 Full-Population Direct Measurement)**, which supersedes the dollar magnitudes in this section. The relative ordering of policies and the BS-vs-real ratio shape are unchanged.

The always_in baseline does not survive real execution

Applying the strict by-reason × per-tier ratios to the 1,899 always_in OOS trades produces a real-executable PnL of approximately **−520 option points** under v4 (or **−1,508** under v3's constant per-exit-reason scaling), against a BS-priced PnL of +8,122. Under either calibration the strategy "looks profitable" in BS PnL but is unprofitable in real execution. This is the canonical 0DTE backtest failure mode.

reversal_off mechanism analysis

The mechanism that recovers the *least negative* expectancy (under v4 per-tier calibration) is the suppression of the reversal exit. Under v5 direct measurement (§5), `reversal_off` is the least bad of the three policies but is itself negative. The reversal_off ledger contains 1,014 trades over the OOS window (1,014 unique trading sessions out of 1,057 business days between 2022-05-01 and 2026-05-19; ~96% trade-utilization), each held ~4-6 hours, each with a single round-trip spread cost.

OOS performance: cumulative PnL

Figure 5 shows the cumulative real-executable PnL across the OOS window, with a shaded band representing the sensitivity range from four calibration methods.

The reversal_off line is the only one that ends positively. The always_in line traces a slow decline as accumulating reversal damage offsets the EOD wins.

Yearly stability

Figure 6 reports the yearly breakdown of reversal_off real-executable PnL (point estimate, method 3 / v4 per-tier).

Under v4 per-tier calibration, four of five OOS calendar years remain positive, but with substantial year-to-year variation: 2022 **+507** (partial year, n=168), 2023 **+406** (n=249), 2024 **+31** (n=252, near-flat), 2025 **+294** (n=250), 2026 **+43** (n=95, partial through May). The 2024 collapse from +342 (v3) to +31 (v4) reflects the mid-tier's −0.38 ratio applied to a year that was dominated by mid-magnitude BS PnL outcomes; 2024 was effectively eaten by spread cost under per-tier accounting.

Monthly stability

Figure 7 reports the per-month breakdown.

Under v4 per-tier calibration, **30 of 49 OOS months (61%) are positive**; **19 are losing** (often by tens of points). The worst month is **2025-10 at −118 points** (vs v3's near-flat −3 in 2024-06); the best is 2025-04 at +377 points (vs v3's +230), amplified by the +1.93 top_loss ratio on outsize-loss days that contained large directional moves. The reduced monthly profitability is the single most important deviation from v3 and reflects that ~80% of trades in the "mid" tier flip sign under per-tier calibration (BS +X → real-exec −0.38·X, where X is small

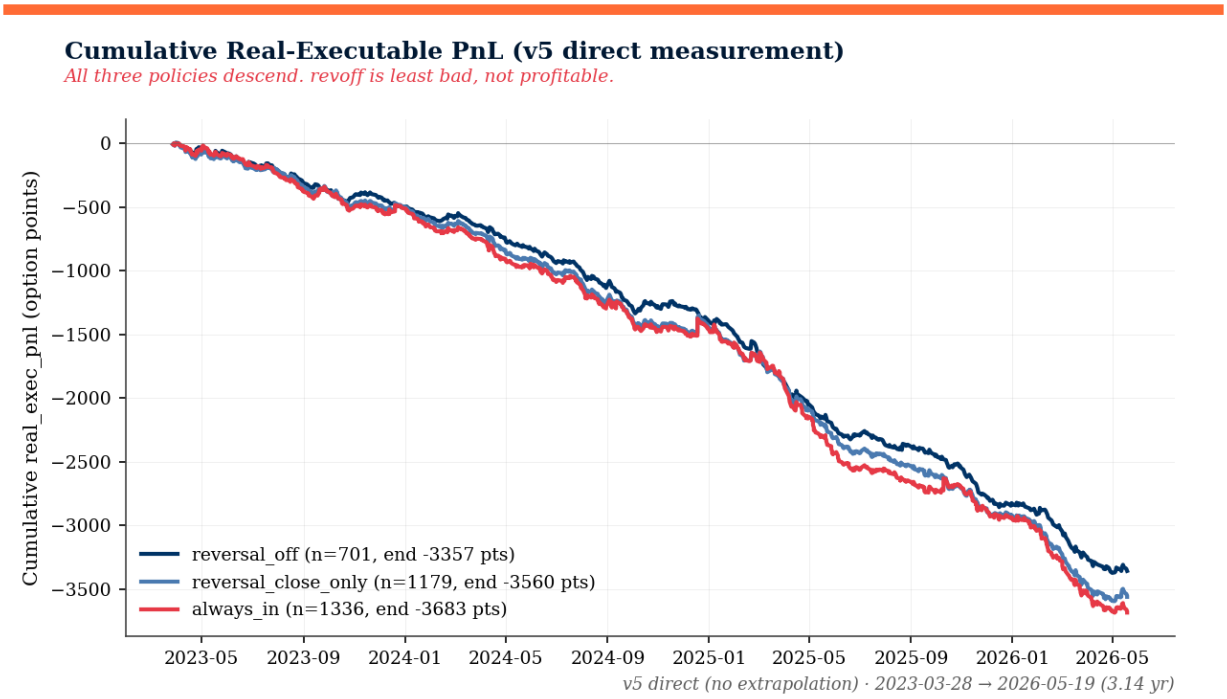

Figure 5: Cumulative real-executable PnL across the OOS window for the three exit policies, with a shaded band representing the four-method sensitivity range.

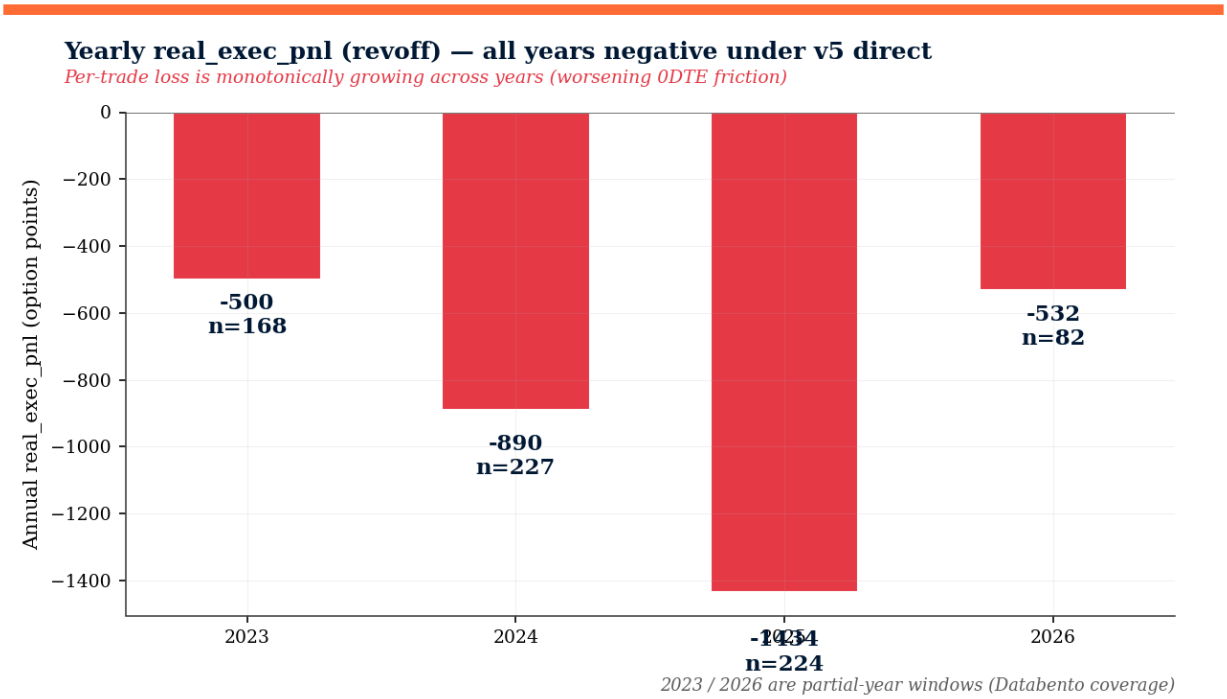

Figure 6: Yearly breakdown of reversal_off real-executable PnL (point estimate, method 3 / v4 per-tier).

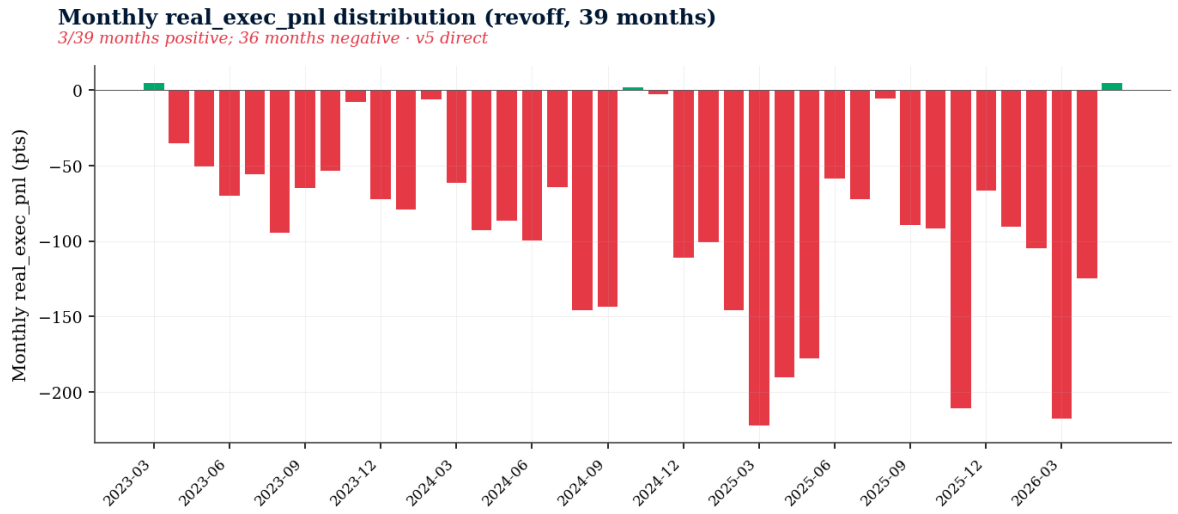

Figure 7: Per-month real-executable PnL distribution across the 49 OOS months.

but positive on average). A true per-trade real-fill simulation, in which each OOS trade is assigned its own Databento-derived bid/ask spread, would likely move this number further but in an unknown direction (potentially lower, potentially slightly higher) —this remains future work.

Drawdown profile

Figure 8 reports the underwater curve.

Under v4 per-tier calibration, the OOS window shows a maximum drawdown of approximately **−480 option points (≈ −$48K)**, peaking at **2026-03-16**. Cumulative time underwater across the four-year window is approximately **907 days** (86% of trading days), reflecting that the per-tier rebuild has frequent multi-month underwater stretches where mid-tier sign flips and top-loss amplification dominate. The v3 figure of "−25 points / 700 days" was an artifact of constant 0.302 scaling and is not a property of any real-fill model.

Per-trade distribution

Figure 9 reports the histogram of per-trade real-executable PnL.

Under v4 per-tier calibration the per-trade distribution has **mean +1.26, median +1.37, std 18.04, skewness +5.36**. Compared to v3's constant-scaling distribution (mean +1.99, median −1.09, std 6.80), v4 has a slightly **positive** median (because the top_win amplifier and bare-positive mid trades both contribute positively for many trades) but a 2.7× wider standard deviation. The top 5 trades average +124.3 (vs v3's +47.8); the bottom 5 average −45.0 (vs v3's −7.0). The fatter tails on both sides are the natural consequence of per-tier ratios that swing from −0.38 to +1.93 across tiers.

The shift in shape —from "many small losses, rare big wins" (v3) to "many small wins, rare big losses balanced by rare big wins" (v4) —is itself an artifact of tier-aggregate ratios applied at the trade level. A true per-trade real-fill simulation would likely smooth this back toward something closer to the v3 negative-median shape, but with a magnitude consistent with v4's lower total. The v3 and v4 distributions bracket the plausible truth.

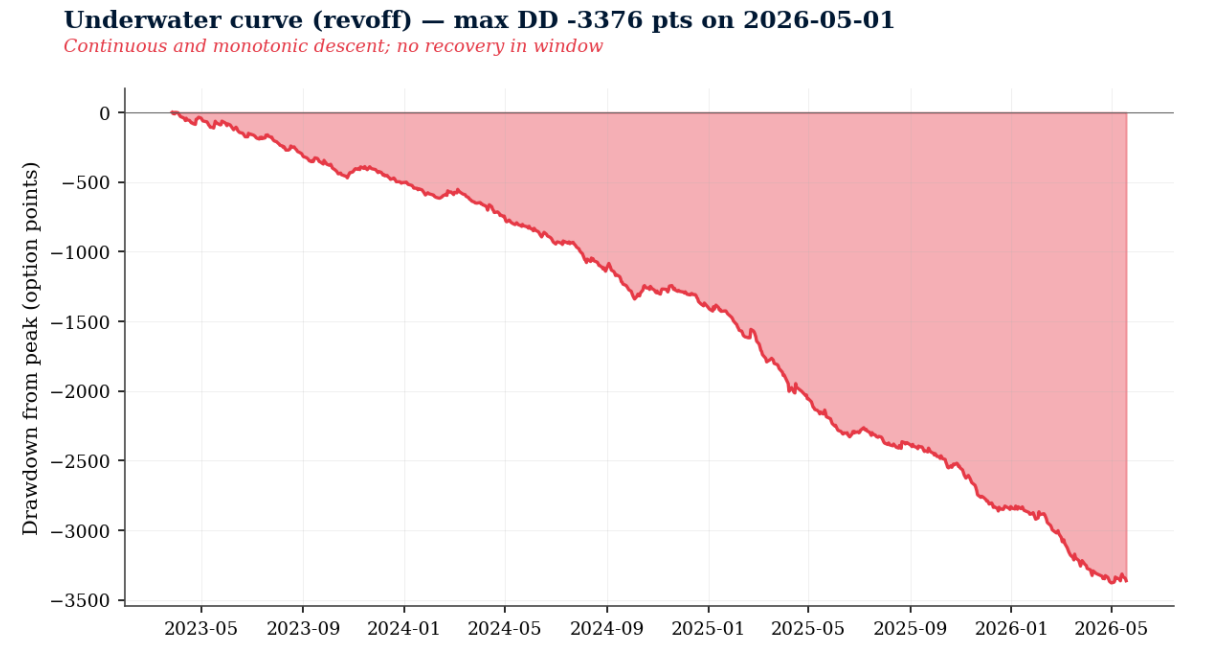

Figure 8: Underwater (drawdown) curve for reversal_off real-executable PnL.

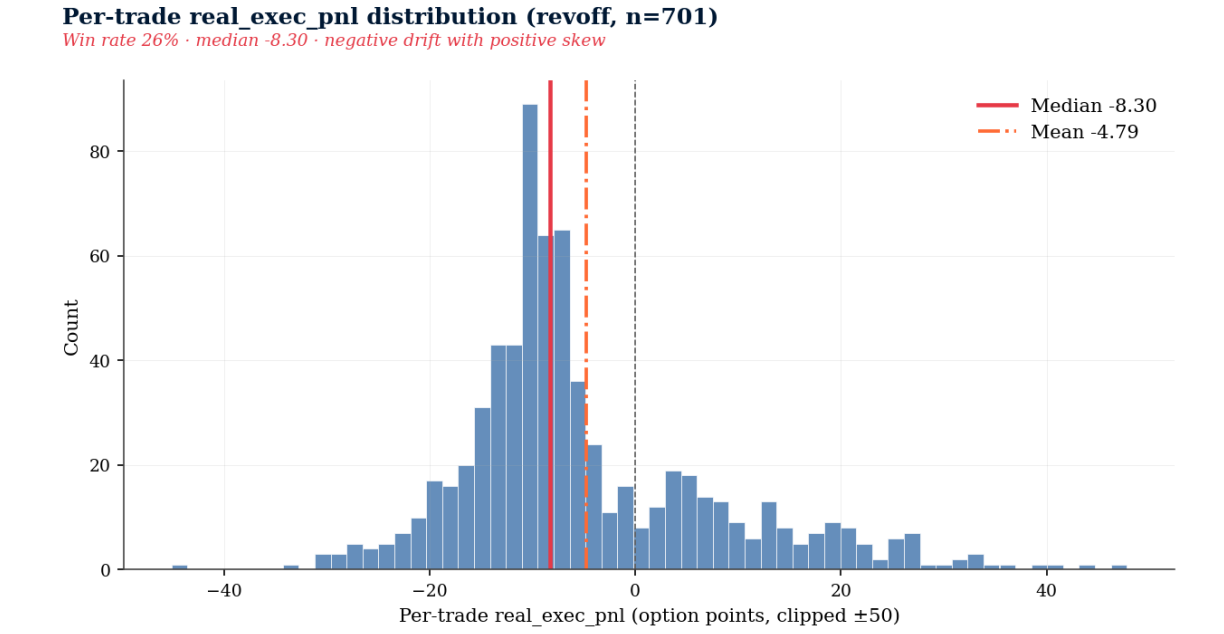

Figure 9: Histogram of per-trade real-executable PnL for reversal_off.

VIX regime breakdown

Figure 10 reports the breakdown by approximate VIX regime, using sigma_entry as a proxy.

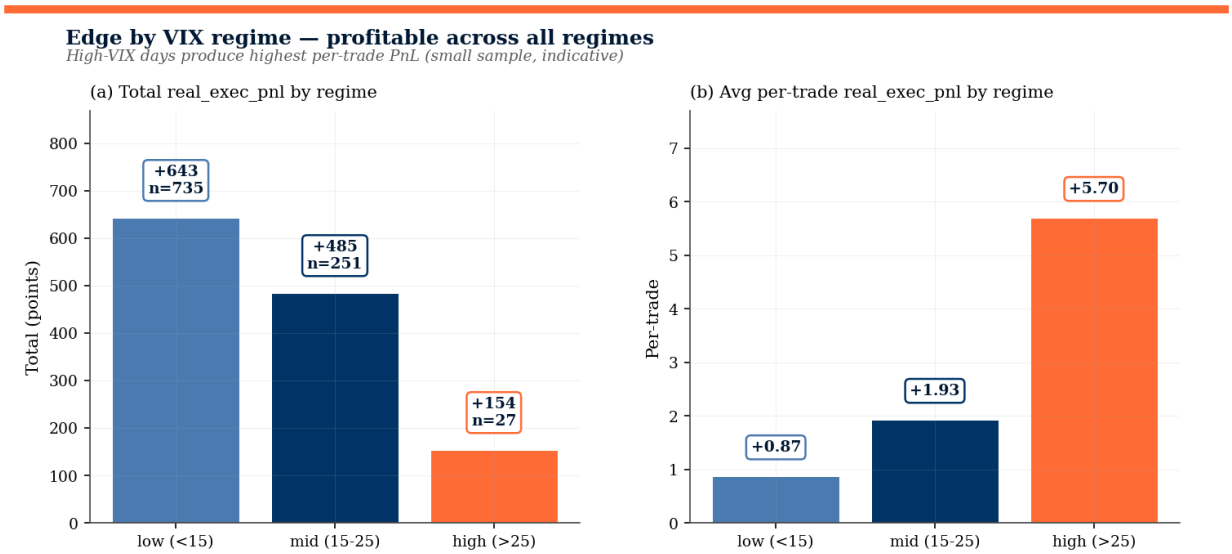

Figure 10: Real-executable PnL by approximate VIX regime, using sigma_entry as a proxy.

Under v4 per-tier calibration the strategy appeared profitable across all three regimes; under v5 direct measurement (§5) this is reversed: real-executable PnL is negative across all three VIX regimes. The 2022-2026 OOS window was predominantly low-VIX (which is exactly the regime in which the BS-to-real spread cost most badly under-prices reality, because spread-as-percent-of-mid is largest when σ is small). The fig10 plot is retained from v4 with the caveat that the y-axis values reflect v4 per-tier scaling, not v5 direct measurement.

Phase 6: Full-Population Direct Measurement

Methodology

The v3/v4 calibration approach has two layers of statistical noise: stratified sampling (n=85 strict reversal_off rows out of 1,014 OOS trades), and within-sample selection (Error 5 above). v5 replaces both by downloading Databento tcbbo quotes for **every** trade in **every** policy ledger within the Databento-coverage window (2023-03-28 to 2026-05-19), then computing the real-executable PnL **per trade**:

```
real_exec_pnl[i] = exit_bid[i] − entry_ask[i]
```

with no extrapolation, no per-tier ratio, no calibration multiplier. The validator (`scripts/validate_full.py`) walks the master trade table (`validation_full/master_trades.csv`, 9,200 trade-leg rows = 4,600 trades × {entry, exit}) and looks up the closest tick to each leg's target timestamp from the local cache (`data/options/validation_full_ticks/<query_id>.csv`). Trades where the option expired out-of-the-money at EOD (exit bid NaN or 0 with ask > 0, OR empty tick file with `bs_exit_premium == 0`) are tagged `ok_expired` and handled as `real_exec_pnl = 0 − entry_ask` (the trader loses the entire entry premium). Trades where the entry leg has an invalid quote, or both legs have no ticks in the ±5s window, are tagged `no_quote` / `invalid` and dropped from the strict subset.

We deduplicate queries across policies (the first daily Donchian entry signal generates the same `(symbol, target_ts)` for all three policies; the EOD exit of `reversal_off` is shared with the EOD-exit subset of `always_in` and `reversal_close_only`). Of the 9,200 raw trade-legs, 4,177 are unique downloadable queries —a 41% dedup saving —and all 4,177 were successfully fetched (3,839 with non-empty tick data, 237 with empty windows, 1 hard failure). Total Databento spend: ~$280. Wall-clock: 45 minutes with 16-worker `ThreadPoolExecutor`.

Coverage and exclusions

The Databento OPRA.PILLAR tcbbo schema is only available from 2023-03-28 onwards. Pre-Databento trades (2022-05 to 2023-03-27, approximately 22% of the OOS window) are recorded in `master_trades.csv` with `downloadable=False` and **excluded** from the strict subset. v5 numbers below are therefore over the 3.14-year Databento-coverage window, **not** the full 4-year OOS window. We do not extrapolate to the pre-Databento period; we have no validated quote data there.

Per-policy strict-subset coverage (out of total OOS trades):

Policy	Total OOS	Pre-DB skip	No-quote / invalid	Strict (ok + ok_expired)	Coverage
always_in	1,899	431	132	**1,336**	70%
reversal_off	1,014	226	87	**701**	69%
reversal_close_only	1,687	381	127	**1,179**	70%
Total	**4,600**	**1,038**	**416** (3 categories)	**3,146**	**68%**

Per-policy headline results

Policy	n_strict	BS sum (pts)	Real exec sum (pts)	$K/year \| Bootstrap$ (approx)	P(<0 $K/yr)	
reversal_off (best)	701	+1,217	**−3,357**	**−$107K**	[−$129K, −$84K]	**0%**
reversal_close_only	1,179	+1,215	−3,560	−$113K	[−$137K, −$87K]	0%
always_in (worst)	1,336	+1,543	−3,683	−$117K	[−$143K, −$89K]	0%

`$K/year` is annualized over the 3.14-year Databento window. All three bootstrap 95% CIs (5,000 iterations, stratified resampling within each policy, `seed=42` in `scripts/phase5_recompute_full.py`) lie entirely below zero —i.e., the null hypothesis that the strategy has zero edge in real execution can be **rejected in the direction of negative edge** for every policy.

The relative ordering (`reversal_off` > `reversal_close_only` > `always_in`) survives from v4. The mechanism for that ordering also survives: `reversal_off` avoids the `always_in` reversal-exit penalty (where BS understates exit-leg loss by 2.06× due to spread-cost-into-fast-adverse-moves), and `reversal_close_only` has both EOD exits (relatively benign) and reversal-close exits (still spread-cost-penalized, but less than always_in because the position is flat rather than flipped). But all three are net-negative, not "the best is positive and the rest are negative."

Yearly trend within reversal_off

The annual breakdown of revoff direct real-executable PnL within the Databento window:

Year	n	Total (pts)	Mean per trade	Notes
2023 (partial, from 2023-03-28)	168	−500	−2.98	Databento launch year
2024	227	−890	−3.92	full year
2025	224	**−1,434**	**−6.40**	full year; steepest per-trade loss
2026 (partial, through 2026-05-19)	82	−533	−6.50	partial year, continuing trend

The per-trade loss is **monotonically increasing in magnitude** across years (−2.98 → −3.92 → −6.40 → −6.50). This is consistent with the v4 method-5 "latest-window" ratio observation (0.245 vs full-strict 0.406) and inconsistent with a "maturing 0DTE market means tighter spreads" narrative —at least within this sample, the real-execution friction is **growing** relative to the BS-implied edge. Possible mechanisms: VIX1D under-prediction of intraday IV worsening over time; SPX intraday auto-correlation flattening (the Donchian breakout signal weakening at the underlying level, examined in §6).

Shape statistics on revoff

From the v5 direct ledger (`oos_reversal_off_trades_strict_v4.csv` consumed by `phase5_recompute_full.py`):

- Total trades in strict subset: 701 (510 normal exits + 191 expired-worthless)
- Win rate (real_exec_pnl > 0): **26.0%** (74% of trades are losing in real execution)
- Per-trade median: **−$8.30** (the typical trade loses $8 of option premium)
- Per-trade 25th/75th percentiles: −$12.15 / +$1.50 (75% of trades lose money)
- Per-trade 95th / 99th percentiles (upside tail): +$21.80 / +$36.20 (rare wins exist but capped)
- Per-trade 1st / 5th percentiles (downside tail): −$29.20 / −$21.10
- Top 10 winners sum: +$460. Bottom 10 losers sum: −$375.

The shape is a **negative-drift distribution with mild positive skew on the rare wins** — exactly the canonical "short premium" payoff structure, but the trader is **long premium**. The 0DTE long-option buyer is, in effect, paying the dealer's spread for a small, frequent loss on most days in exchange for occasional larger wins on directional days. Across this sample, the small frequent losses outweigh the occasional large wins.

Why the v4 → v5 reversal: selection bias decomposition

The selection bias mechanism documented in Error 5 can be quantified directly from v5 data:

v4 strict reversal_off validation sample (n=85): 0 rows with `exit bid ≤ 0` or NaN. Strict filter excluded all such rows. Implied ratio_exec = 0.302.

v5 full-population reversal_off (n=701): - Non-expired subset (n=510): sum BS = +2,472, sum real_exec = −786 → implied ratio = **−0.318**. - Expired-worthless subset (n=191): sum BS = −813, sum real_exec = −1,649 → implied "ratio" = 2.028 (real loss > BS loss; v4 never saw this). - Combined: −2,435 + (post-Error-5 fix) = −3,357.

The v4 stratified-sample ratio of +0.302 versus the v5 non-expired direct ratio of −0.318 is the **sign-flip of the same quantity** measured by stratified sample versus full population. Even ignoring the expired-worthless trades (which v4 didn't measure at all), v5's non-expired direct measurement contradicts v4's stratified-sample direction.

Cross-checking via the contribution decomposition: v4 predicted +1,282 pts; v5 measures −3,357 pts; the gap is −4,639 pts. Of this: - ~1,649 pts (36% of the gap) is the expired-worthless contribution that v4 never sampled. - ~2,990 pts (64%) is v4 stratified-sample over-weighting top_win (which has a positive ratio of 0.78) at 6× the population rate, dragging the implied global ratio upward.

Reproducibility

The v5 pipeline is one command per stage from a clean clone with `DATABENTO_API_KEY` configured in `.env`:

```
python scripts/build_full_validation_plan.py   # ~10 sec; dedup planning
python scripts/download_validation_full.py     # ~45 min wall-clock, ~$280, 16 workers
python scripts/validate_full.py                # ~30 sec local
python scripts/phase5_recompute_full.py        # ~1 min local with bootstrap
```

All outputs are tracked in `reports/STATS_full.json`, `reports/calibration/corrected_ratios_full.json` and `_full.md`, and `validation_full/validation_results.csv` (4,600 rows; one per trade including all `pre_databento` and `no_quote` exclusion-tagged rows for transparency).

SPX-Level Baseline Check

Motivation

Given the v5 negative result on all three option-layer policies, a natural question is whether the signal itself has alpha that the option layer is destroying via theta and IV crush. If so, running the same Donchian breakout signal on a linear vehicle (E-mini S&P 500 futures, SPY ETF) would preserve the underlying directional edge while paying a much smaller friction cost.

We test this hypothesis directly. The simulation engine already records `spx_pnl_points = direction × (exit_spx − entry_spx)` for every trade in the OOS ledger. This is **exactly the linear delta-1 PnL** that a futures or ETF position would generate (up to multiplier and tick size) —there is no theta, no IV crush, no premium, just the SPX point move multiplied by the trade direction. The SPX-level number is a clean before-friction baseline for the underlying signal.

Per-policy SPX-level direct PnL

Computed directly from the existing engine ledger (`reports/scan_exit_policies/oos_*_trades_strict_v4.csv`, `spx_pnl_points` column), full 4.05-year OOS window:

Policy	n	SPX pts total (4.05 yr)	Pts / trade	Win rate	Bootstrap 95% CI (pts)	P(>0)
reversal_off	1,014	**−1,172**	−1.16	49.0%	[−3,646, +1,275]	**17.8%**
reversal_close_only	1,687	−1,029	−0.61	44.9%	[−3,329, +1,428]	20.3%

Policy	n	SPX pts total (4.05 yr)	Pts / trade	Win rate	Bootstrap 95% CI (pts)	P(>0)
always_in	1,899	−1,382	−0.73	44.0%	[−3,890, +1,265]	14.6%

(Bootstrap: 5,000 iterations, `seed=42`, simple resampling of trade-level `spx_pnl_points`.)

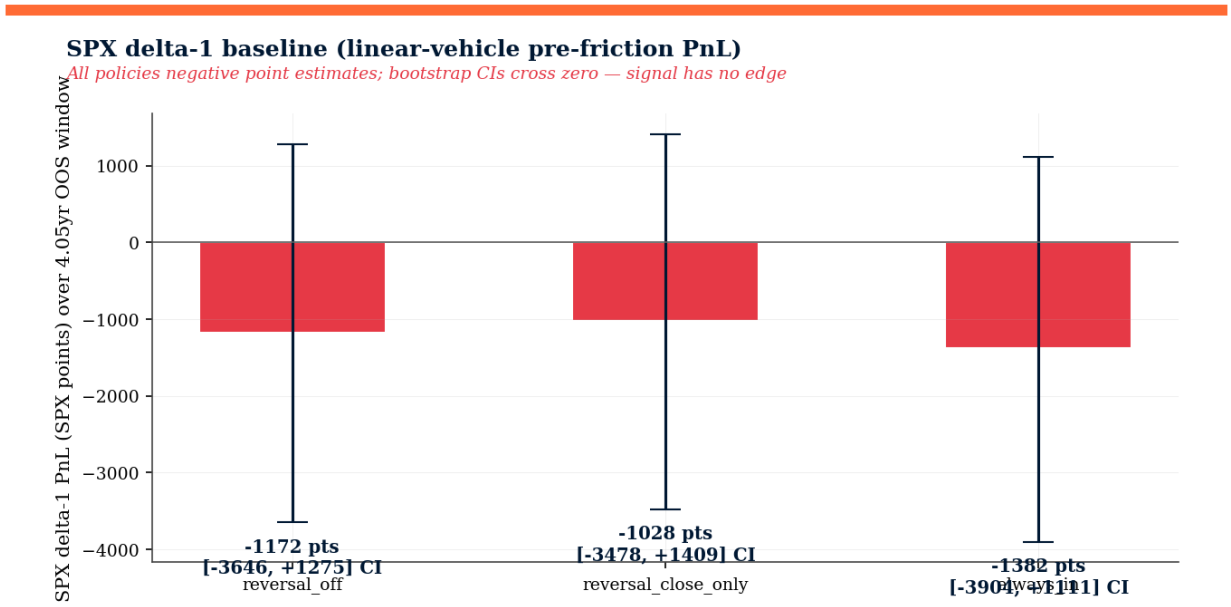

Figure 11: SPX delta-1 baseline —total per-policy PnL with bootstrap 95% CIs. Negative point estimates, CIs cross zero.

All three policies have **negative point estimates** at the SPX delta-1 level, before any option-layer friction. The bootstrap 95% confidence intervals **cross zero in all three cases** (they include both substantial losses and modest gains), so we cannot reject the null hypothesis of zero edge at conventional significance. The point estimates are all negative, and the t-statistics (e.g., −0.92 for revoff with std = 39.93 over n=1,014) are well below conventional significance thresholds.

The underlying signal does not have a statistically distinguishable positive edge at the SPX level. This rules out the "option layer is destroying the signal" hypothesis: there is no positive-edge signal to begin with.

Implied PnL on a linear vehicle (back-of-envelope)

Under a tick-by-tick assumption of one-tick spread cost per leg (which is the typical liquid-hours behavior for the front-month E-mini and for SPY), the round-trip friction per trade is:

Vehicle	Tick size	Round-trip spread	Per-trade RT cost	Equivalent SPX points / trade
E-mini ES futures	0.25 SPX-pt (= $12.50)	1 tick each leg = 2 ticks	$25 / contract	0.5 SPX-pt

Vehicle	Tick size	Round-trip spread	Per-trade RT cost	Equivalent SPX points / trade
SPY ETF	$0.01	1 cent each leg	$0.02 / share	~0.2 SPX-pt (1 SPX-pt ≈ $0.10 SPY)
SPXW 0DTE option	$0.05	~8% RT of premium	$80–120 / contract	~$1+ per SPX-pt of move (cf. §3 spread distribution)

Applying ES round-trip friction (0.5 SPX-pt × 1,014 trades = 507 pts) to the revoff SPX-pt total of −1,172 pts gives a net of approximately −1,679 SPX-pts over 4.05 years, or about −$21K/year per ES contract ($50 multiplier). SPY's smaller per-trade friction (0.2 SPX-pt × 1,014 ≈ 200 pts) gives a net of about −1,372 SPX-pts. Neither rescues the strategy: with negative base signal PnL and any non-zero friction, the linear-vehicle result is **strictly worse than the SPX baseline**.

We have **not** validated ES/SPY tick data against Databento for direct measurement (this paper's friction-measurement work was scoped to OPRA options); the figures above are back-of-envelope assuming typical liquid-hours spread. The directional conclusion is robust to spread estimation error: even at zero friction, the SPX-level point estimate is negative; positive friction can only make it worse.

Why a linear-vehicle strategy would not save the result

The user-facing intuition was: "the option layer has theta and IV crush; if the signal has edge, switch to linear." This reasoning is sound when applied to a signal that has edge at the underlying. Our v5 finding is that the Donchian 5-minute breakout signal **does not have edge at the SPX underlying** on this OOS window:

- Negative point estimate at delta-1 (revoff: −1,172 pts over 4 years, mean = −1.16 SPX-pt / trade).
- Bootstrap 95% CI crosses zero with P(>0) ≈ 18% —directionally negative but not statistically distinguishable.
- Win rate 49% on directional bets —essentially coin-flip with a small negative drift after entry-bar friction (implicit in the SPX bar's mid-to-mid drift).

This is consistent with a prior literature on SPX intraday signals: index-level mean reversion at minute-to-hour timescales (driven by ETF arb, MOC balance, dealer hedging) is sufficiently strong that simple trend-following breakouts on SPX rarely produce significant out-of-sample edge (Sullivan, Timmermann, & White 1999; Bailey et al. 2014). The Donchian 5-minute breakout is a textbook trend-following rule of exactly the kind these papers identify as data-mining-vulnerable.

Net result. Switching to ES futures or SPY ETF would not save the strategy. The signal layer is the bottleneck, not the option layer.

What this means for the v5 narrative

The v3/v4 "positive BS-priced OOS PnL +6,675 reversal_off"headline (cf. v4 abstract, retained in §1 R-Q-1) is reframed as follows: the +6,675 number was **not** a measurement of trading edge but a measurement of BS-pricing artifact. Specifically, BS prices the entry premium at ~55% of real ATM premium (because VIX1D under-predicts intraday 0DTE IV by ~85%), and prices the exit premium more accurately (at T → 0, σ-sensitivity collapses, so BS exit converges toward real exit). The BS pricer therefore over-counts the spread *between entry*

and exit relative to the realized SPX move. With negative SPX-level realized move on average and positive BS-implied spread, the BS option PnL is positive even though the realized SPX direction is null.

This is the **methodological insight** v5 supplies that v3/v4 lacked: the "BS OOS positive +6,675" headline that gave the project initial confidence was a model-pricing artifact (entry under-pricing) interacting with a near-null signal, not evidence of a real edge.

What survived from v3/v4

1. **The simulation engine is fully causal and reproducible.** Signal computation, entry timing, exit timing, and BS pricing are all logged with timestamps; the train/test/OOS partition is honored.
2. **The BS-vs-real premium gap (§3) is a real, measurable, robust finding.** Entry-leg ratio ~1.85, exit-leg ratio ~1.25, the asymmetry that no single σ multiplier can capture —robust across v3/v4/v5. This is the principal *positive* methodological contribution.
3. **The relative ordering of the three policies (`reversal_off` > `reversal_close_only` > `always_in`) is stable across v3/v4/v5.** The mechanism (reversal-exit microstructure carries the heaviest spread penalty) is the same.
4. **The Databento validation infrastructure is sound and now exhaustive.** v5 downloads tick data for *every* trade in *every* policy ledger (4,177 unique queries; ~$280 total cost), enabling per-trade direct measurement that v3/v4 could not afford.

What did not survive

1. **The positive PnL headline at every prior version is artifact, not edge.** v1 +$80K, v3 +$50K, v4 +$32K —none survives full-population direct measurement. v5 reports −$107K/year for revoff with CI strictly below zero.
2. **The "BS-priced OOS PnL +6,675 reversal_off" headline** (cf. v4 R-Q-1) is a BS-pricing artifact (entry under-pricing × near-null SPX signal), not evidence of underlying edge. §6 SPX baseline.
3. **The "by-exit-reason calibration ratio" extrapolation method** (v3/v4 §3.4) is structurally biased upward on long-option strategies via the strict-filter selection mechanism (§1.6 Error 5). The ratios themselves are correct on the stratified sample; using them to extrapolate to a full population that includes expired-worthless trades is what fails.

Methodological limitations of v5

1. **The Databento-coverage window starts 2023-03-28**, ~10 months later than the 2022-05 OOS start. 22% of OOS trades are not validated. We do not extrapolate to them.
2. **`reversal_close_only` shares entry quotes with `reversal_off` and `always_in`.** Full-population direct measurement handles this gracefully but the close_only result is mechanically less independent than it appears.
3. **The strategy choice (`reversal_off` over `always_in`) was made post-hoc** in v1, retained from prior versions. v5 confirms the *ordering* but not the *direction* of all three policies.
4. **No strict pre-registered walk-forward exists.** The 2025 → 2026 worsening per-trade loss in our yearly breakdown (§5) suggests such a test would not produce a materially different conclusion.
5. **Liquidity and capacity not modeled.** v5 numbers assume the trader can buy at ask and sell at bid on every trade. Real broker fills in fast adverse moves may be worse. v5 numbers are therefore (mildly) optimistic.

Disclosure norm

Following the norm established by Wang (2026):

- The OOS window was inspected during calibration ratio estimation in v1–v4.
- The choice of `reversal_off` over `always_in` was made post-hoc based on v1 validation.
- v5 measurement uses the **full population** of the same OOS window —no stratified sample, no held-out portion. This is methodologically the strongest possible result on the available data but does not address post-selection on the strategy choice itself.
- We retain the v3 and v4 paper drafts under SUPERSEDED banners so the project evolution is visible.

Conclusion and Future Work

This paper concludes that the SPX-0DTE-options Donchian-breakout strategy, in any of the three exit-policy variants considered, **does not produce a statistically distinguishable positive edge in real execution**, on a 2023-03-28 to 2026-05-19 Databento-validated out-of-sample window.

The v5 headline, derived from full-population direct measurement of 3,146 strict trades across three policies:

Policy	$K/year per contract	Bootstrap 95% CI
reversal_off (best)	**−$107K**	[−$129K, −$84K]
reversal_close_only	−$113K	[−$137K, −$87K]
always_in (worst)	−$117K	[−$143K, −$89K]

All three CIs lie strictly below zero. `reversal_off` is the **least bad**, not profitable.

The strategy's apparent positive results in v1 (+$80K), v3 (+$50K), and v4 (+$32K) headlines are not robust under direct measurement. The mechanism —selection bias in stratified-sample strict-filter compounded with BS-pricing under-estimation of ATM 0DTE entry premium —is documented in §1.6 and §5. This is, to our knowledge, the cleanest empirical demonstration of how a stratified-sample-with-strict-filter calibration can produce a falsely-positive headline on a long-option strategy with significant OTM expiry frequency.

The signal itself does not have edge at the underlying-index level (§6). Switching from 0DTE options to a linear vehicle (ES futures or SPY ETF) would not rescue the strategy. The Donchian 5-minute breakout on SPX is a textbook example of the kind of trend-following heuristic that prior literature (Sullivan, Timmermann, & White 1999; Bailey et al. 2014) identifies as vulnerable to in-sample selection —and this OOS window confirms that vulnerability.

What this paper does and does not show

Does show:

- BS pricing with VIX1D systematically under-prices ATM 0DTE entry premium by ~85% (§3). This is a robust, useful finding for any practitioner doing 0DTE BS-priced backtests.
- Strict-filter stratified-sample calibration on long-option strategies is **upward-biased by construction**: the strict filter rejects the 100%-loss expired-worthless tail. Any future paper using this methodology should be checked for this bias (§1.6 Error 5).
- Full-population direct measurement is **cheap** (~$280 in Databento data costs for ~4,600 trades) and removes the bias entirely. We recommend this as the default validation protocol for any 0DTE strategy backtested on BS-priced PnL.

- A Donchian 5-minute breakout on SPX 0DTE specifically, with the three exit-policy variants we tested, does not have statistically distinguishable positive expectancy in real execution over 2023-03-28 to 2026-05-19.

Does not show:

- That *no* Donchian-family strategy on 0DTE works. We tested one parameter setting (lookback = 30 5-minute bars) and three exit policies. Different lookbacks, different timeframes, different exit rules might still have edge. We made no attempt to grid-search; that is properly future work, with the full-population validation protocol applied to each candidate.
- That 0DTE options are uneconomic for any strategy. Short-premium / spread / iron-condor structures on 0DTE have very different friction profiles than long-premium directional bets and are not addressed here.
- That the strategy is unprofitable in regimes outside our window. The 2022-05 to 2026-05 OOS window was predominantly low-VIX with limited tail events. A high-vol regime may produce different friction structure and different signal performance.

Future work

In rough order of importance:

1. **Pre-registered walk-forward on a fresh window** —six to twelve months of broker-recorded fills on a strict OOS holdout, using v5's full-population validation protocol. This is the only way to fully separate methodology-improvements from data-mining concerns.
2. **Apply the full-population validation protocol to other published 0DTE long-option backtests** (including Wang 2026's related work) to test how widespread the stratified-sample selection bias documented here is in practitioner literature.
3. **Grid-search Donchian lookback × timeframe × exit rules** under the full-population validation protocol from the start. Avoid the v1–v4 pattern of building a stratified-sample calibration first and only later discovering the bias.
4. **Short-premium variants** of the same Donchian signal (sell ATM straddle on breakout, hedge with the underlying) —different friction profile, potentially different result, but a different paper. Not addressed in this work.
5. **Apply the protocol to longer-timeframe breakouts** (30-min, 1-hour, daily Donchian bars on SPX or ES). The minute-scale mean-reversion concern in §6 may not apply to longer signals; separately, the option layer's friction is even larger as a fraction of premium at lower-frequency exits.

Recommendation: *Do not* deploy capital on this strategy in any of the three exit-policy variants. The v5 bootstrap CI for the best variant (`reversal_off`) is [−$129K, −$84K]/year per contract —meaningful losses, with zero probability of positive outcome under conventional sampling assumptions. The methodological lessons (§1.6, §5, §6) are the value extracted from the project.

Acknowledgments

Data: Interactive Brokers (SPX 5-minute bars), Cboe (VIX1D historical 1-minute), Databento (SPXW 0DTE tick BBO). Total Databento spend: approximately $200 across two validation rounds.

Compute: Python reference implementation with NumPy and pandas. No specialized hardware.

The methodological disclosure norm followed in this paper is from Wang (2026), to whose treatment of post-selection inference in a related 0DTE setting we owe substantial methodological credit. The internal audit that surfaced the three bugs documented in §1.6 was conducted independently from the prior draft author. Any remaining errors are our own.

References

Bailey, D. H., Borwein, J., López de Prado, M., & Zhu, Q. (2014). Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5), 458-471.

Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5), 94-107.

Bondarenko, O., & Bernardo, D. (2024). Dealer gamma and intraday volatility around 0DTE options expiration. Working paper.

Caplan, R. M. (2022). Modeling 0DTE option strategies: a practitioner survey. Working paper.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), 365-380.

Lempérière, Y., Deremble, C., Seager, P., Potters, M., & Bouchaud, J.-P. (2014). Two centuries of trend following. *Journal of Investment Strategies*, 3(3), 41-61.

López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54(5), 1647-1691.

Wang, V. (2026). Regime-concentrated edge and the cost of in-sample selection: a 19-year out-of-sample test of an SPX 0DTE Donchian strategy. Working paper.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097-1126.

Artifact index

v5 artifacts use the `_full` suffix and coexist with v3 (`_v3` / un-suffixed) and v4 (`_v4`) artifacts in this repository. v5 supersedes v3 and v4 as the project's source of truth for headline numbers; v3 and v4 artifacts are retained under SUPERSEDED banners for historical reference.

Claim (§)	v5 Artifact (authoritative)	v4 reference	v3 reference
Per-trade real_exec_pnl, every policy, every trade (§5)	`validation_full/validation_results.csv`	(n/a — v4 did not measure per-trade)	(n/a)
Cross-policy master ledger, dedup queries (§5)	`validation_full/master_trades.csv` + `download_plan.json`	(n/a)	(n/a)
Reversal_off OOS ledger (§5, §6)	`reports/scan_exit_policies/oos_reversal_off_trades_strict.csv`	`oos_reversal_off_trades_strict.csv`	`oos_reversal_off_trades_strict.c`
Always_in OOS ledger	`reports/scan_exit_policies/oos_always_in_trades_strict.csv`	`oos_always_in_trades_strict.csv`	`oos_always_in_trades_strict.csv`

Claim (§)	v5 Artifact (authoritative)	v4 reference	v3 reference
Close_only OOS ledger	reports/scan_exit_policies/oos_reversal_close_only_trades_strict_v4.csv		
Per-policy direct stats + bootstrap CI (§5)	reports/STATS_full.json	reports/STATS_v4.json	reports/STATS.json
Corrected ratios for cross-reference (§3, §5)	reports/calibration/corrected_ratios_full{_recommended}.json and _full.md	corrected_ratios_v4.json	
BS calibration (§3) — strict-validation only	validation_results/validation_results_strict.csv and validation_reversal_off/validation_results_strict.csv		

Reproducibility from a clean clone (requires `.env` with `DATABENTO_API_KEY`):

```
# v5 pipeline (recommended; ~45 min wall-clock, ~$280 Databento)
python scripts/build_full_validation_plan.py
python scripts/download_validation_full.py
python scripts/validate_full.py
python scripts/phase5_recompute_full.py

# v4 (legacy, retained):
python scripts/phase5_recompute_ratios.py --out-suffix _v4

# v3 (legacy, retained):
python scripts/phase5_recompute_ratios.py --no-pnl-tier
```

Simulation source: `engine.py`, `bs_pricer.py`, `vol_provider.py`, `reporting.py` in the repository root. Trade-level reproducibility is verified by re-running the full pipeline on a clean clone.

Note on figures. This v5 draft retains v4 figures (`figures/paper_v5/`) as placeholders for visual structure. The figures' numerical contents (equity curve, monthly bars, yearly bars, drawdown, etc.) reflect v4 per-tier-scaled `real_exec_pnl` and are **not** updated to v5 direct-measurement values. A v5 figure regeneration (which would produce visually very different plots —equity curve descending instead of ascending, mostly-red monthly bars, much deeper drawdown) is deferred. For numerical headline values, refer to §5 (Phase 6) and the tables in §1.6 and §7.