

Prices That Never Existed: A Universe-Scale Measurement of Phantom-Print Contamination in Equity Backtests

Daniel Gatto*¹

¹Universidade Paulista (UNIP)

July 2026

Abstract

A printed trade reports that a transaction happened, not that an order could have traded there: odd lots, off-exchange prints, and condition-coded trades are unreachable by a marketable order or excluded from the official high, low, and last. A bar built from all prints therefore sits above the executable high and below the executable low, so every stop, target, and breakout is scored against prices that never existed. From one trades feed we build naive and executable bars by print-time condition codes, across the US equity top-1000 (2007–2026, 1.46 billion regular-hours bars); three results follow. First, 42.3% of bars carry a phantom-affected extreme, rising to 90% by 2026, led by odd-lot prints (76% of the range inflation). Second, the headline is selection fidelity, not profit: across 7,232 strategies in an edge-free universe, in-sample selection on executable bars predicts the executable out-of-sample ranking better than on naive bars (rank-IC 0.9987 versus 0.9948; gap +0.0039, $p < 0.001$ under an effective- N cluster bootstrap). Third, a CME futures negative control near zero and the 2013 odd-lot-tape-onset natural experiment isolate the effect to equity microstructure. A formal companion traces the mechanism partly to persistence, so a strategy can pass naive validation in and out of sample yet die on execution (the deployment trap); a known-answer planted-drift experiment on a 60-name panel finds the clean-selected book higher at every tested drift strength, consistent with a deployment flip it suggests rather than directly observes. The claim is one-sided: cleaning restores selection fidelity rather than manufacturing an edge.

Keywords: market microstructure, odd lots, off-exchange trading, sale condition codes, OHLC bars, backtest overfitting, selection fidelity, walk-forward.

JEL classification: G11, G14, C12, C15.

*Correspondence: daniel@daru.finance. Replication code: <https://github.com/DaruFinance/phantom-print-bars>.

Contents

1	Introduction	3
2	Related Work and the Gap	5
3	Institutional Background	8
4	Data	9
5	Methodology	11
5.1	Bar construction: naive versus executable	11
5.2	The phantom-range metric	12
5.3	Intrabar fill model	12
5.4	The selection experiment	12
5.5	Statistical machinery	14
5.6	The range is sign-definite, but the backtest distortion is two-sided	14
6	Results I: Prevalence and Magnitude of Untradeable Extremes	15
7	Results II: Decomposition of Range Inflation by Print Source	18
8	Results III: Selection Fidelity, the Headline	18
9	Results IV: Intrabar Ambiguity, Tick-Resolved	22
10	Robustness	24
11	Negative Control: CME Futures	26
12	Corollary Nulls	27
13	A Measurement-Error Model of Phantom-Print Overfitting	28
13.1	Primitives and the executable-edge null	28
13.2	Strategies and the measurement channels	29
13.3	The finite-moment factor representation	30
13.4	The deployment-fidelity gap	31
13.5	Same-tape fidelity and the deployment trap	32
13.6	Selection value, genuine edges, and the Gaussian benchmark	34
13.7	Dispersion, identification, and empirical scope	36
13.8	Reported planted-drift experiment	39
14	Discussion	40
15	Conclusion and Limitations	42
A	Condition-Code to OHLC-Eligibility Mapping	44
B	Reproducibility and Data Availability	44
C	Results Inventory	45
D	Figure Provenance	46
E	Pre-Registration Hash	46

1 Introduction

A trader reads a chart. The bar's high is a number printed on the tape, and the natural assumption is that the number is a price someone paid and therefore a price the trader could have paid. That assumption is wrong often enough to overturn a backtest. The US consolidated tape carries a print whenever a transaction is reported, and a transaction is reported under many conditions that do not correspond to a price a marketable order could have reached at that instant: the trade was an odd lot too small to set a protected quote, it was executed off-exchange and reported to a FINRA facility seconds late and possibly out of sequence, or it carried a sale condition (average price, prior reference price, derivatively priced, extended hours) saying the price was set by a formula or an earlier time rather than the current market. The exchanges know this. The SIP plan specifications mark, for each sale condition, whether the trade may update the official high, low, and last, field by field, and the rule is that if a condition says a trade may not update a given field, suppression wins for that field. The official OHLC the exchanges publish already applies these per-field eligibility rules. The defect this paper measures lives one layer down, in the raw trades feed that most practitioners actually consume and aggregate themselves: a bar builder that runs a maximum and a minimum over raw prices reintroduces every print the exchanges deliberately suppressed.

The reason this matters for a backtest, and not merely for a descriptive statistic, is that the contamination is structural. The bar high is almost always set by a buyer-initiated print and the low by a seller-initiated print, so even a perfectly clean trade-only range is wider than the executable range by about a spread (Corwin and Schultz [11]). Phantom prints widen the wicks further, and the range can only widen, never narrow (Proposition 1). But the effect on a strategy is not one-directional. A breakout that triggers off a phantom high books a winner the market never offered, which flatters it; a long strategy whose protective stop sits just above a phantom-depressed low records a spurious stop-out, a loss the market never offered, which deflates it. Across the full universe the inflating case is the more common one (it flatters the in-sample Sharpe of most strategies), but a material minority are deflated. So phantom prints do not uniformly flatter a backtest; they push different strategies in different directions by strategy-specific amounts, and in doing so they scramble the in-sample performance ranking. A scrambled ranking is one that no longer predicts which strategies will work out of sample. The question this paper answers is how much that ranking corruption costs, and whether selecting strategies on cleaned bars is more faithful to live performance than selecting on the contaminated ones.

Contributions. This paper contributes four parts. First, a measurement of the prevalence and magnitude of untradeable prints in OHLC extremes: how often, and by how much in basis points, the naive high or low is set by a print the exchanges themselves do not count, resolved by name and time of day. Second, a diagnostic decomposition of the range inflation by print source (an order-dependent attribution, not a causal identification), separating odd-lot, off-exchange “D”, other ineligible-condition, and residual bid-ask-bounce contributions, with the bounce baseline anchored to the Corwin-Schultz and EDGE estimators so the phantom-print share is identified net of the spread component that survives any cleaning. Third, a structural result, Proposition 1: the executable print set is a subset of all prints up to a rare late-tape edge case, so the naive bar's high is weakly above the executable high and its low weakly below, and the phantom range inflation is sign-definite. We are careful to scope this to the range only: the resulting backtest

distortion is two-sided, and we measure it to inflate the in-sample Sharpe of the majority of strategies (85% of the corpus) through phantom breakouts while deflating a minority (15%) through spurious stop-outs, a two-sided scramble of the in-sample ranking rather than a uniform shift in either direction. Fourth, the centerpiece: a strategy-selection experiment across 7,232 strategies spanning twelve structural families (reduced to ≈ 20 mutually uncorrelated cross-symbol portfolios), in which the only thing varied is the bar set used to select and optimize each strategy in-sample, and the out-of-sample truth is always the executable bar set, because a live order can only fill at an executable price. The finding is not profit. The universe is edge-free by construction (intraday strategies net of real costs leave nothing profitable out-of-sample), which is the cleanest setting in which to isolate the contamination, because any apparent in-sample edge is by construction false. The pre-registered primary endpoint is selection fidelity: how faithfully an in-sample ranking of strategies predicts their executable out-of-sample ranking. Selecting on clean executable bars (Arm C) predicts the executable out-of-sample ranking with rank-IC 0.9987, higher than selecting on the naive bars practitioners use (Arm N, 0.9948), a gap of +0.0039 that is positive and significant under an effective- N cluster bootstrap ($p < 0.001$) and rises to +0.020 on the mutually uncorrelated subset. Both arms exceed 0.99 because a strategy’s persistent cost-and-turnover signature dominates the ranking in an edge-free universe, so the quantity that matters is the naive-versus-clean gap, not the absolute level. An internal-validity control (Arm N scored on its own naive out-of-sample, rank-IC 0.991) confirms the naive backtest looks self-consistent precisely because it is contaminated identically in and out of sample. This is the paper’s central claim, and it is a null on profit: cleaning restores fidelity, it does not manufacture a real edge. A battery of controls supports it: the dose-response is within-equity specificity evidence, a tick-resolved intrabar measurement shows the tie-breaking channel is small, a random-print-drop placebo makes selection no better than naive (ruling out a “fewer prints” explanation), and a two-arm CME futures negative control running the identical twelve-family universe finds the futures phantom analog rare and the fidelity gap statistically zero, isolating the contamination to equity microstructure.

Scope. We scope every headline to what is tested. The universe is the full US equity top-1000 names by dollar volume, point-in-time and including delisted names, with prevalence and decomposition measured over the full 2007–2026 tape and the strategy experiment confined to the odd-lot era 2014–2026; the strategy family is the structural shapes a path-dependent intraday backtest exercises (breakout and channel, stop-limit bracket, intrabar take-profit and stop-loss, mean reversion); the cleaning is the print-time condition-code filter described in Section 5. The window begins in 2014 because odd-lot executions were absent from the consolidated tape before December 2013 (Section 3), so the phantom-print mechanism operates only in the odd-lot era; the 2013 onset itself is our cleanest natural experiment (Section 10). The claim is a null on profit: removing phantom prints raises the fidelity of an in-sample selection and removes the differential harm of naive selection, and it does so because the phantom prints were distorting in-sample performance in both directions, scrambling the ranking, not because cleaning adds information. Cleaning cannot manufacture an edge in a strategy that has none; what it does is stop a contaminated backtest from mis-ranking strategies, which is exactly what we observe here, where nothing is profitable and naive selection still does measurable damage. The complementary positive claim, that clean bars let a genuinely profitable strategy survive while naive bars degrade it, requires a universe that contains profitable

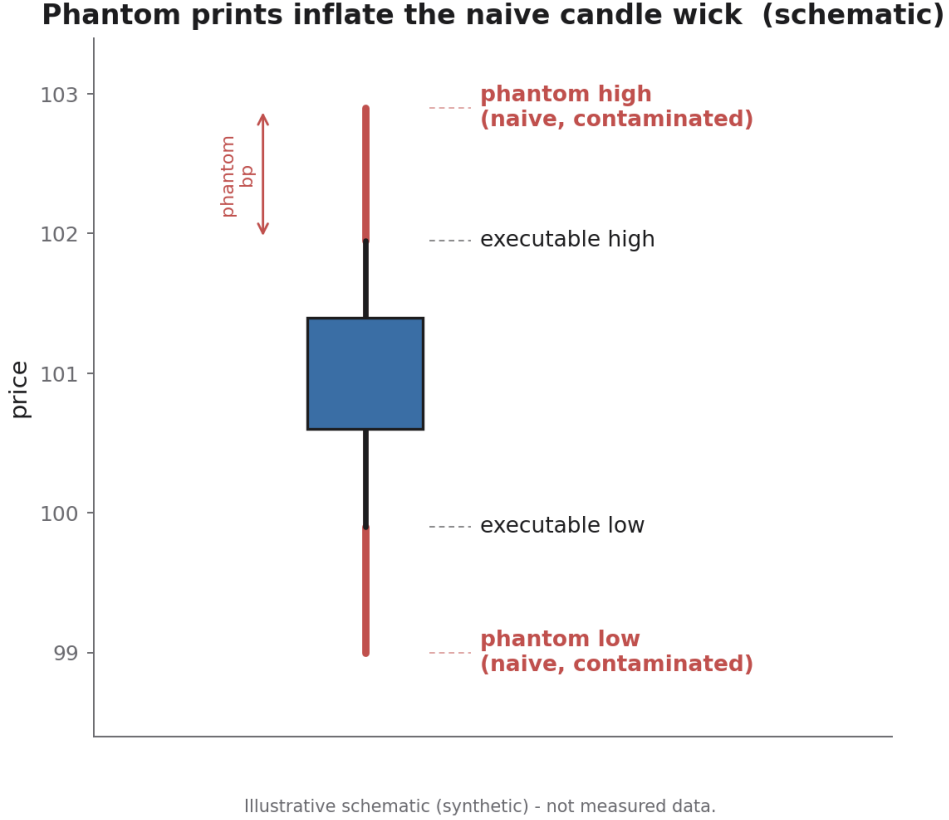


Figure 1: The phantom candle. A naive one-minute bar (left) takes its high and low over every print on the tape, including an odd-lot print above the executable high and an off-exchange/late print below the executable low; the executable bar (right) keeps only prints reachable at or inside the NBBO. The wicks a path-dependent strategy trades against shrink to the prices a marketable order could actually have reached.

strategies. The full top-1000 universe over 2014–2026 is edge-free as well: net of real costs no naturally occurring strategy has a positive executable out-of-sample Sharpe. We therefore establish that claim two ways, by proof and by a known-answer experiment that plants a controllable edge in the tape. Theorem 3 shows that naive selection forfeits a strictly positive amount of executable out-of-sample performance whether or not an edge is present, Corollary 2 gives the band of genuine edge on which the difference is between a deployable book and a losing one, and Remark 3 evaluates the resulting penalty at our measured contamination parameters; a planted-drift experiment (Section 13.8) then finds the clean-selected book higher at every tested drift strength, consistent with that band, although the opposite-sign interval is interpolated from the reported settings rather than directly observed. Sections 11 and 12 give the controls and corollary nulls that hold us to this reading.

2 Related Work and the Gap

Four literatures meet at this paper, and it fills a stated white space between them.

Odd-lot information content. O’Hara, Yao, and Ye [20] established, using proprietary Nasdaq message data matched to TAQ, that before the December 2013 tape change a

median stock had roughly a fifth to a quarter of its trades missing as odd lots, more than 60 percent for high-priced names, and that those missing odd lots carried roughly a third of price discovery because informed and high-frequency traders deliberately shred orders into sub-round-lot clips. Bartlett, McCrary, and O’Hara [8] supplied the quote-side analog, documenting an inside market of better-priced odd-lot quotations the NBBO ignores because protected quotes must be round lots. These papers prove odd lots are informative and that excluding them biases volume and order-flow measures. They are about the information in odd lots, not about what odd-lot prints do to a backtested trading rule.

TAQ data-quality bias. Holden and Jacobsen [14] are the closest prior art and the most important contrast. Comparing cheap Monthly TAQ against expensive Daily TAQ, they show timestamp granularity and quote handling change measured effective spreads by a large margin and produce many more nonpositive spreads, distorting trade-location and price-impact estimates. Their result is that the data product materially changes a measurement. Ours is the missing companion: the data product, specifically whether the bar excludes phantom prints, materially changes whether a backtested strategy survives out of sample. They bias liquidity estimation; we bias strategy selection and optimization, which is the practitioner-facing failure mode.

OHLC range bias and microstructure noise. Corwin and Schultz [11] make the on-point observation that the bar high is almost always a buyer-initiated trade and the low a seller-initiated trade, so the high-low range is inflated by the spread, and they turn this into a spread estimator. The trade-direction sign underlying that observation is the classic Lee and Ready [15] inference, and the spread the range is inflated by is, in the informed-trading account, compensation for adverse selection (Glosten and Milgrom [12]). Ardia, Guidotti, and Kroencke [2] generalize it to all four OHLC prices in their EDGE estimator and show the bias is worst when trading is infrequent. Roll [21] gives the bid-ask-bounce mechanism. The realized-variance literature (Zhang, Mykland, and Ait-Sahalia [24]; Ait-Sahalia, Mykland, and Zhang [1]; Bandi and Russell [6]; Hansen and Lunde [13]; Barndorff-Nielsen, Hansen, Lunde, and Shephard [7]) shows that naive variance from the finest prints diverges to the noise variance rather than converging to true volatility. We use these as the residual baseline: the spread component is the part of range inflation that survives any cleaning, so the phantom-print contribution must be measured net of it. Our claim is not that we discovered the OHLC range is biased; it is that a specific, removable, condition-coded subset of that bias drives backtest non-replication.

Intrabar fills and backtest overfitting. Maier-Paape and Platen [18] and Low, Maier-Paape, and Platen [17] formalize intrabar fill ambiguity: with only OHLC, when a stop and a target both sit inside one bar the engine cannot know which was hit first, the case is not uniquely decidable, and engines that resolve it optimistically overstate performance. Bailey, Borwein, Lopez de Prado, and Zhu [3] prove backtest overfitting is near-certain once enough configurations are tried, and the deflated Sharpe ratio and PBO via CSCV (Bailey and Lopez de Prado [5]; Bailey, Borwein, Lopez de Prado, and Zhu [4]) deflate a winning backtest by its effective number of trials. We treat both as machinery our experiment must respect, not as the contribution: the intrabar ambiguity is a separate channel we measure on ticks in Section 9, and the overfitting deflation is applied at the assembly step in Section 5.5. Independently, Mesfin [19] runs a systematic falsification of

fourteen OHLCV intraday signal families on a single futures instrument and finds none clear realistic costs, a strategy-corpus null adjacent to our edge-free finding but on one instrument and without the naive-versus-executable bar contrast that is our subject.

Practitioner tooling and vendor infrastructure. The phenomenon is not unknown to practitioners; parts of it are encoded in market-data infrastructure. The SIP high/low-eligibility rules themselves already exclude odd-lot and many condition-coded prints from the official OHLC (Nasdaq Equity Technical Update 2013-33 and the CTA and UTP output specifications; see Section 3), and data vendors expose the per-condition eligibility as a field. On the cleaning side, practitioner tooling names the artifact directly: the Concretum Group’s `concretum_tests` library detects “phantom highs/lows” in one-minute OHLCV data alongside stale and missing bars [10]. What none of these does is measure what the contamination costs the strategy-selection step of a backtest, decompose it by print source, or contrast naive against executable selection across an uncorrelated strategy corpus. We take the naming and the eligibility rules as given and supply the measurement.

The gap. The literature has measured odd-lot information content, odd-lot quotes inside the NBBO, TAQ liquidity-measurement bias, and OHLC spread bias. No prior work has measured how naive OHLC-bar construction degrades the strategy-selection step of a backtest, decomposed that contamination by print type, and shown across a corpus of uncorrelated strategies that selecting on executable bars is more faithful to live performance than selecting on naive bars, with a futures control isolating the effect to equity microstructure. That is the program of this paper.

Addressable universe. The question reaches the full US equity tape, the top-1000 names by dollar volume over the odd-lot era 2014 to 2026, plus the matched CME futures used as the cross-market control, and this paper runs on exactly that universe: 998 names and 1.46 billion regular-hours bars over 2007–2026 for the descriptive prevalence and decomposition, with the strategy experiment on the odd-lot era 2014–2026 (3,158 trading days). The framework was developed and frozen on a stratified eight-name three-year development panel, fixing every design choice before the full universe was analysed; every headline below is the full-universe number. We establish the effect where the universe is edge-free, so the result is the fidelity gain of clean selection rather than a profitable strategy: net of real costs no strategy in the full corpus has a positive executable out-of-sample Sharpe. The complementary positive result, that executable bars let a genuinely profitable strategy survive while naive bars degrade it, together with the realized dollar penalty, therefore has no naturally occurring example in this edge-free universe. We establish it by proof (Theorem 3, Corollary 2), evaluate the penalty at the measured contamination parameters (Remark 3), and probe it with a labelled known-answer planted-drift experiment (Section 13.8). This is a program-complete result: the genuine-edge case a naturally occurring edge-free tape cannot supply is addressed by proof and, suggestively, by that experiment, which finds the clean-selected book higher at every tested strength without directly observing the opposite-sign interval or identifying a dollar loss.

3 Institutional Background

This section gives the minimum the result needs. It is the dossier's institutional material compressed to what the result needs.

Protected quotes, round lots, and odd lots. Under Regulation NMS (SEC Release 34-51808, 2005) [25], Rule 611(a) requires each trading center to establish, maintain, and enforce written policies and procedures reasonably designed to prevent trade-throughs of protected quotations, and Rule 611(b) enumerates the exceptions to that duty (intermarket sweep orders, benchmark and derivatively-priced trades, stopped orders, flickering quotations displayed for less than one second, and quotations of a trading center in a state of material delay), which are the same sale conditions Table 1 records. A protected quotation is defined in Rule 600(b), not in Rule 611: it must be an automated (immediately executable), top-of-book quotation of at least a round lot, so an odd-lot quotation is not a protected quotation and an execution at or through an odd-lot price is not a trade-through. The practical consequence is that the official market can show one bid and offer all minute while a trade prints outside it, and that is not an error: it is an odd lot, an off-exchange print, or a condition-coded trade, none of which had to respect the NBBO. The round lot itself was conventionally 100 shares until the 2020 Market Data Infrastructure Rule (Release 34-90610) tiered it by price band, with the 2024 Reg NMS amendments (Release 34-101070) accelerating the change and adding a best-odd-lot-order identifier to the SIP.

Odd lots on the tape. Odd-lot executions were entirely absent from the consolidated tape and from TAQ until December 9, 2013, when CTA and UTP plan amendments (SEC Release 34-70794; Nasdaq Equity Technical Update 2013-33) began reporting them with a new trade modifier "I". Critically, odd-lot prints are flagged volume-eligible only: they update consolidated volume but do not update the official Last, High, or Low, and do not move LULD bands. The SIP therefore already excludes odd lots from official OHLC. A raw trades feed still contains the odd-lot prints, and a naive bar builder picks them up. By October 2021, odd lots were a majority of all trades by count, driven by high nominal prices, where a 100-share round lot is prohibitively large, and by the retail and fractional-share boom.

Off-exchange and the TRF. Roughly 40 to 50 percent of US share volume executes off-exchange, in ATSS and through wholesaler internalization, none of which display a quote. When these execute, the broker-dealer reports the trade to a FINRA Trade Reporting Facility (FINRA Rule 6300A/6300B and 6380A/6380B), and on the tape the print carries the "D" exchange code. These prints can sit outside the NBBO for three legitimate reasons. They are not protected quotes, so a dark or internalized fill can print at, inside, or outside the NBBO without a trade-through. Reporting carries latency: FINRA rules require a report as soon as practicable but no later than 10 seconds after execution, with later reports flagged late and out of sequence, so a TRF print can hit the tape seconds after the actual execution at a price that was the market then but not now. And benchmark trades (VWAP, average price, prior reference price, qualified contingent trades) carry a price set by a formula or an earlier time. The economic driver is payment for order flow: commission-free brokers route marketable retail orders to wholesalers who internalize them at a sub-penny improvement over the NBBO and pay the broker for the

flow (CRS IF11800; the December 2022 SEC reform package). The order never touches a lit exchange, and the print it leaves was not accessible to a competing marketable order.

Sale conditions and OHLC eligibility. Every trade message on CTS and UTDF carries sale condition codes, and the SIP plan specifications (CTS Pillar Input/Output Specification; the UTP Sale Condition Matrix) define, for each condition, whether the trade updates High, Low, Last, and Volume, separately for the consolidated and the participant tapes. The governing rule is suppression-wins: if a trade carries multiple conditions and any one says do not update, the trade does not update. Table 1 gives the material subset. Good vendors encode this. Algoseek normalizes all SIP conditions into a 32-bit bitmask (bit 31 `tOddLot`, bit 10 `tFormT`, bit 20 `tAveragePrice`, bit 25 `tPriorReferencePrice`, bit 9 `tDerivativelyPriced`, bit 5 `tIntermarketSweep`, bit 14 `tOutOfSequence`, bit 29 `tTradeThroughExempt`, among others) and exposes a TRADE-NB (trade-at-NBBO) event recommended for backtesting. The full mapping is Appendix A.

4 Data

The source is the Algoseek US equity tape: tick-level trades and the trade-only minute-bar products. Algoseek collects the full SIP (Tapes A, B, and C) and normalizes the per-trade sale-condition modifiers into the 32-bit bitmask of Section 3, and it exposes the TRADE-NB event. Odd lots are flagged for the source decomposition by a fixed quantity-under-100-shares test on the trade tape, the round-lot convention in force across most of the sample, not by the price-tiered round-lot definition (Section 10 states the one place this matters). The results below are computed on the full US equity top-1000 universe by dollar volume, 998 names. The tape spans 2007 to 2026, 1,464,795,700 regular-trading-hours executable bars, and the prevalence and decomposition of Sections 6–7 are measured over that full span, which lets the descriptive surface carry the odd-lot era transition directly. The selection experiment of Section 8 is confined to the odd-lot era, 2014-01-02 to 2026-07-24, 3,158 trading days, because the phantom-print mechanism requires odd lots on the tape and they were absent before December 2013 (Section 3). All design choices were pre-specified on a stratified eight-name development panel spanning microstructure regimes (NVDA, AAPL, TSLA high-turnover odd-lot-heavy; SPY, QQQ, SQQQ off-exchange-heavy ETFs; BAC, F low-priced) over 2021–2023, before the full universe was analysed; the numbers reported below are the full-universe numbers throughout.

Feed-confound discipline. The naive bars and the executable bars are built from the same underlying trades feed by filtering condition codes, never by comparing two differently-reconstructed Algoseek products. This is deliberate. A separate Algoseek feed-commensurability finding from prior work showed that two differently-reconstructed products of the same tape can differ materially on identical dates, structurally, from the reconstruction itself rather than from any economic effect. Building both bar sets from one feed by filtering makes the naive and clean bars commensurable by construction, so any measured difference is attributable to the filter and not to the reconstruction.

Point-in-time and survivorship. The universe is point-in-time and includes delisted names, so a strategy is selected and evaluated only on the names that were tradeable on each date. Survivorship leakage would itself inflate apparent edge and confound the

Table 1: Sale-condition codes and OHLC eligibility (material subset). Each row gives the condition, the Algoseek bitmask flag, and whether the trade is eligible to update the official last sale and, separately, the official high and low; the two can differ by field, and the complete per-field table is Appendix A. Source: CTS Pillar and UTP specifications; FINRA Trade Report Modifiers; Algoseek Trade-Only guide.

Condition	Flag	Last?	High/Low?	Note
Regular-way trade	(none)	Yes	Yes	The eligible base case.
Intermarket sweep (ISO)	tIntermarketSweep	Yes	Yes	Eligible despite multi-price execution.
Opening/closing auction trade (“O”/“6”)	tOfficialOpen / tOfficialClose	Yes	Yes	Auction execution; eligible for last, high, and low.
Official open/close message (“Q”/“M”)	(indicative)	No	No	Indicative price message, not a trade; the naive bar builder must exclude it.
Odd lot (“I”)	tOddLot	No	No	Volume-eligible only since 2013-12-09; the bars flag odd lots by a fixed under-100-share test (Section 4).
Average price (“W”)	tAveragePrice	No	No	Computed average, not a market print.
Prior reference (“P”)	tPriorReferencePrice	No	Yes	Updates high and low, not last.
Derivatively / benchmark	tDerivativelyPriced	No	Yes	VWAP/benchmark price; updates high and low, not last.
Form-T / ext. hours (“T”)	tFormT	No	No	Pre/post market.
Out of sequence (“Z”)	tOutOfSequence	No	Yes	Late/out-of-sequence report; updates high and low, not last.
Off-exchange “D”, regular	code “D”, no condition	Yes	Yes	Last-sale eligible. The executable bar keeps the at-or-inside-NBBO subset and excludes only the out-of-NBBO subset.
Off-exchange “D”, flagged	“D” with .P/.W/late/QCT/Form-T	No	per flag	The suppressed subset: late, prior-reference, average-price, benchmark, or extended-hours “D” prints.

contamination measurement, so the universe construction is documented and the delisting handling is stated in Appendix B.

Granularity. Prevalence (Section 6), decomposition (Section 7), and the strategy experiment (Section 8) run at minute-bar granularity, both bar variants built from the same trades. The intrabar measurement (Section 9) runs at tick granularity on nine symbol-days. Prevalence and decomposition are aggregated over the full 1.46-billion-bar universe; the strategy experiment pools all 998 names into cross-symbol strategy portfolios, so its unit of analysis is the strategy, not the symbol-strategy pair.

Futures control data. The cross-market control (Section 11) uses the Algoseek CME futures tape: five front-month roots (ES, NQ, CL, GC, ZN) over 1,250 contract-days in 2023, a single central limit order book per root, and no SIP sale-condition bitmask or off-exchange venue. The tape does carry away-from-market prints (block, EFRP, TAS, implied or spread-leg), reported as non-aggressor trades; the clean filter keeps only aggressor-typed trades, which is the futures analog of the equity phantom-print cleaning. The identical bar construction and selection pipeline run on this tape unchanged, which is what makes it a genuine two-arm negative control.

5 Methodology

The methodology is stated once, up front. Sections 6 through 12 report results against it and do not restate it.

5.1 Bar construction: naive versus executable

A *naive* bar takes the maximum and minimum over all prints in the interval, which is what a practitioner who ignores condition codes builds. The *executable* bar is built from Algoseek’s TRADE-NB event, defined in the spec as a trade executed *at or within* the prevailing NBBO. Its extremes are therefore the prices a marketable order could actually have reached at the instant they printed. The TRADE-NB filter is a price condition, not a venue condition: it retains every reachable print at or inside the NBBO, including price-improved inside-NBBO fills and midpoint off-exchange “D” prints (which are reachable, indeed generated, by marketable orders), and it excludes only out-of-NBBO prints, the ones a marketable order could not have reached. This makes the executable bar a different object from the official OHLC the exchanges publish, neither a subset nor a superset of it. Official OHLC suppresses odd lots, average price “.W”, Form-T “.T”, cash, next-day, and qualified-contingent prints from the high and low, but sale-condition eligibility is per field rather than per trade: prior reference “.P”, derivatively or benchmark priced, and out-of-sequence prints are not last-sale eligible yet do update the official high and low. Official OHLC also retains regular-way off-exchange “D” prints that printed outside the NBBO, which are last-sale eligible and do update official high, low, and last even though a marketable order could not have reached them; the TRADE-NB bar excludes exactly those out-of-NBBO prints while keeping the reachable inside-NBBO “D” prints, so it is not the same as consuming official OHLC fields. Every filter uses only print-time information, the condition code and the contemporaneous NBBO carried with the trade, so the cleaning is causal and carries no look-ahead. A print is assigned to its minute bar

by the single timestamp the trades feed carries, and the naive and executable bars use that identical assignment, so the two arms differ only in the condition-code and NBBO filter and never in bar membership. For off-exchange “D” prints that timestamp is the FINRA report time, which can trail execution by up to ten seconds (Section 3); a “D” print executed at the prior market but reported once the NBBO has moved is then classified against the reported-time NBBO, which is a classification error rather than look-ahead and applies identically to both arms.

5.2 The phantom-range metric

For each bar, the inflation is the naive range minus the executable range, reported in basis points and in ticks. Each inflated extreme is attributed to the specific condition that set it: if the naive high is set by an odd-lot print and the executable high by a regular-way print one tick below, the gap is attributed to the odd-lot condition. We aggregate to per-name and universe distributions and report the full distribution, not the mean, because the contamination is concentrated and a mean understates the tail that drives backtest failures.

5.3 Intrabar fill model

On ticks, we resolve the which-hit-first question directly rather than assuming it. For each bar in which a stop and a target both lie inside the bar, the tick sequence determines which was reached first. We report the optimistic-versus-pessimistic gap (the performance difference between assuming the favorable exit first and assuming the adverse exit first) as the Maier-Paape and Platen test battery prescribes, and we report how often the tick-resolved order flips a trade’s outcome relative to the pessimistic convention. This isolates the share of edge inflation that comes from the intrabar path assumption, as distinct from the bar-extreme contamination that is the paper’s focus.

5.4 The selection experiment

This is the payoff. On the edge-free universe it is a measurement of differential harm and selection fidelity, not a commercial profit proof; the profit claim requires a genuine edge, which the edge-free tape lacks and a known-answer planted-drift experiment (Section 13.8) probes (see also Sections 8 and 15).

Corpus. We generate strategies following the structural-versus-tunable convention: a strategy is a structural shape, and the numeric parameters (lookback, threshold, stop and target distances, maximum hold) are in-sample-tunable knobs sampled per walk-forward window, not enumerated as separate strategies. The corpus spans twelve structural families. The first six are a technical-analysis grid (moving-average, oscillator, channel, and bracket shapes), 2,048 specifications per symbol. The other six are structural families ported from a documented strategy generator, 864 specifications each (5,184 in total): compression-breakout (enter on a volatility-compression release), trend-slope (regression-slope trend following), ROC-trend (rate-of-change trend), Donchian-breakout-under-trend (a channel breakout gated by a higher-timeframe trend), higher-timeframe breakout, and session-momentum (an intraday-session momentum gated by a clock). The port reuses the exact signal transforms and gates from the generator and is verified causal: the higher-timeframe trend computed on a series truncated at time t is bit-identical to the full series at the last

common index, the session gate is a pure function of the clock, and the engine fills at $t+1$. The full grid is 7,232 specifications, each evaluated as an equal-weight cross-symbol portfolio over the full 998-name universe, so the unit of analysis is the strategy (one cross-symbol return series), not the symbol-strategy pair. This is the standard trials unit in the overfitting literature (Bailey and Lopez de Prado [5]) and it avoids the 7.2-million-unit inflation that treating each symbol-strategy pair as a trial would introduce. Because these are cross-symbol portfolios of the same specification families, they are far more correlated with one another than per-name strategies would be, so the mutually uncorrelated count is much smaller. Greedy correlation-aware de-duplication (keep a strategy only if its maximum absolute Pearson correlation, on clean out-of-sample daily returns, to the already-selected set is below $\tau = 0.5$, ranking by the mean of the two arms' absolute in-sample Sharpe) reduces the corpus to 20 mutually uncorrelated strategies. We note that τ is this de-correlation threshold (an absolute Pearson correlation cutoff), not Kendall's tau and not the Spearman rank-IC that is the primary endpoint. The de-duplication is applied identically to both arms (the two arms share one corpus), so the Arm C minus Arm N gap is invariant to the corpus-selection step. Because the uncorrelated count is small, we do not rest the headline on the deduped subset alone: the primary estimate (Section 8) uses all 7,232 strategies for the point estimate and an effective- N cluster bootstrap that resamples the ≈ 20 correlation blocks for significance, so the reported p reflects the effective number of independent bets rather than the raw count or any single de-duplication. The dedup-key by τ sweep is the robustness band (Table 6). The corpus count is the number of uncorrelated strategies, not strategies multiplied by walk-forward windows.

Walk-forward and the fixed out-of-sample truth. Selection and optimization use rolling walk-forward with in-sample knob tuning and out-of-sample evaluation, not a single split. The out-of-sample truth is fixed on the executable bars under every arm, because a live order can only fill at an executable price, so the executable out-of-sample is the unbiased proxy for live performance. The only thing that varies across arms is the in-sample bar set.

- **Arm N (status quo).** Optimize and select each strategy's knobs in-sample on naive bars, then evaluate on the executable out-of-sample. This is the faithful counterfactual for the practitioner who backtests on raw bars and then trades live.
- **Arm C (recommended).** Optimize and select in-sample on executable bars, then evaluate on the executable out-of-sample. This aligns the backtest data with the execution reality.
- **Arm N-home (internal-validity control).** Arm N evaluated on its own naive out-of-sample, to show the naive backtest looks internally consistent (in-sample and out-of-sample contaminated identically) yet is externally invalid, ruling out the reading that the gap is just noise.

Metrics. Per strategy, the in-sample-to-out-of-sample replication ratio (executable out-of-sample performance over in-sample performance). Across the corpus, the rank information coefficient (rank-IC), which is the Spearman rank correlation (Spearman ρ) between the in-sample and the executable-out-of-sample per-strategy Sharpe, computed once over the strategy-level cross-symbol corpus (a single coefficient on all 7,232 strategies for the

point estimate, not a per-symbol coefficient then averaged). The selection value: take the top- N by in-sample performance under each arm and compare realized executable-out-of-sample performance of the selected set. And the survival fraction: of strategies in-sample-profitable, what share are executable-out-of-sample-profitable, under each arm. All metrics are net of a fully-specified cost model. Note that the rank-IC (a Spearman rank correlation of performances) is a distinct object from the de-duplication threshold τ of Section 5.4: τ is a correlation threshold on the de-correlation step (keep a strategy only if its maximum absolute Pearson correlation, on clean-out-of-sample daily returns, to the already-selected set is below $\tau = 0.5$), not a rank correlation and not Kendall’s tau, and the two are never the same quantity.

5.5 Statistical machinery

The machinery maps to the paper-bar gates. The null is structural: the executable bar is the control bar, and a bar-shuffle permutation (drop a random equal-count subset of prints) is run through the same pipeline so the effect must be specific to untradeable prints and not to having fewer prints (this is the placebo of Section 12). Inference uses an effective- N cluster bootstrap: because the strategies are cross-symbol portfolios, each of the 7,232 is assigned to its most-correlated de-duplicated representative, forming ≈ 20 correlation blocks, and the headline confidence interval on the rank-IC gap (Arm C minus Arm N) resamples whole blocks, so significance reflects the effective number of independent bets rather than the raw strategy count. The i.i.d.-over-strategies bootstrap is reported only as a diagnostic, to show how far it overstates precision (Section 8). Multiple testing is handled at the assembly step, not per cell: FDR (Benjamini-Hochberg and Romano-Wolf) and the deflated Sharpe ratio deflated by the effective number of trials. PBO via CSCV was planned but is not defined on the de-duplicated corpus, which is too small to split (Section 8). The cost model is a modeled time-of-day half-spread plus a per-fill commission, charged round-trip, and stress-tested at $2\times$, $3\times$, and $5\times$ (Section 10); it does not model gap or impact slippage beyond the half-spread, a limitation we own in Section 15. The analysis plan, including the primary endpoint, is fixed in code before compute and released with its SHA-256 (Appendix E). The same-feed commensurability control of Section 4 is reported. The cleaning is verified not to leak by polluting the walk-forward procedure itself and confirming the pollution is detected, so the no-look-ahead claim is tested at the procedure level and not only at the feature level.

5.6 The range is sign-definite, but the backtest distortion is two-sided

A structural fact governs the bar range, which we state as a proposition. It does not, despite a tempting first reading, imply that the resulting backtest distortion is one-directional.

Proposition 1 (sign-definite range inflation). *Let P be the set of prints in a bar and E the subset reachable at or inside the prevailing NBBO (the TRADE-NB executable set of Section 5.1, which retains inside-NBBO odd-lot and off-exchange prints and excludes only out-of-NBBO prints). If $E \subseteq P$, then the naive high, the maximum price over P , is weakly greater than or equal to the executable high, the maximum over E , and the naive low is weakly less than or equal to the executable low. The phantom range inflation, naive range minus executable range, is therefore ≥ 0 for every bar.*

A maximum over a superset is weakly larger and a minimum over a superset is weakly smaller, which is the whole proof. On the full universe the executable extreme is reconstructed as the union of the executable and the remaining prints, so the naive range is weakly above the executable range on every bar and the sign-definiteness holds exactly (0% of bars carry a negative inflation). This holds under the aligned construction in which the executable and naive bars share the same price units and timestamp-assignment rule; where reporting-time and execution-time assignments differ, set inclusion must be checked bar by bar, since a displaced print can expand the destination bar as well as alter the origin bar, so no one-sided conservativeness claim is made for those cases.

The backtest distortion is two-sided, and we measured it. The proposition is about the range, not about a strategy's profit. A wider range does not flatter a path-dependent backtest one-directionally, the natural but mistaken reading of the proposition. The distortion is two-sided: a phantom-inflated high books a breakout win the market never offered, which flatters a strategy, while a phantom-depressed low triggers a spurious stop-out the market never offered, which deflates one. We compute the signed per-strategy in-sample change (naive minus clean Sharpe) directly across all 7,232 strategies: 85% have their in-sample Sharpe inflated by phantom prints and 15% deflated, with a positive median change of +0.022. So the full-universe distortion is inflation-dominant but genuinely two-sided: phantom breakouts flatter the majority in-sample while spurious stop-outs deflate a material minority.

That two-sidedness is what makes the contamination harmful, and it is the mechanism behind the headline. Because the distortion is two-sided and depends on each strategy's specific stop and target geometry, it does not shift the in-sample performances by a common amount; it scrambles their order. More strategies are pushed up than down, but both happen, by amounts that have nothing to do with executable performance. A scrambled in-sample ranking is precisely a ranking that fails to predict the executable out-of-sample ranking, and that failure is the selection-fidelity gap of Section 8. The harm is ranking corruption, not uniform optimism: a uniform shift would leave the ranking, and therefore selection, intact, whereas the measured two-sided scramble degrades it. A corroborating check: the naive bar produces 1.006 times as many in-sample fills as the clean bar, near-exact turnover parity, so the contamination acts on which prices the fills are scored against, not on how often the strategy trades.

6 Results I: Prevalence and Magnitude of Untradeable Extremes

This section is the descriptive surface area before any strategy is run: how often the naive high or low is set by a print the exchanges do not count, and by how much, over the full 1,464,795,700-bar tape. Because the estimates are population aggregates over the entire universe rather than a small sample, they carry negligible sampling error and we report point values. Over the full 2007–2026 tape, 42.3% of bars carry a phantom-affected extreme. That average hides a sharp regime change at the December 2013 odd-lot tape onset. Before 2014 the phantom-affected share is 2 to 12% and almost none of it is odd-lot, because odd lots were not disseminated to the tape until December 9, 2013, so only the closing weeks of that year carry any odd-lot share and the 2013 annual figure already ticks up from the onset; from 2014 it jumps to 30% and climbs monotonically to 90%

Table 2: Phantom prevalence by year over the full universe, showing the December 2013 odd-lot tape onset. Each row is the phantom-affected-bar fraction, the median range inflation over affected bars, and the odd-lot share of the range inflation, aggregated over all 998 names in that year. Prevalence is near zero and non-odd-lot before 2014, then jumps and climbs across the odd-lot era. Population aggregates over 1.46 billion bars; sampling error is negligible. Source: `prevalence_fixed.json`.

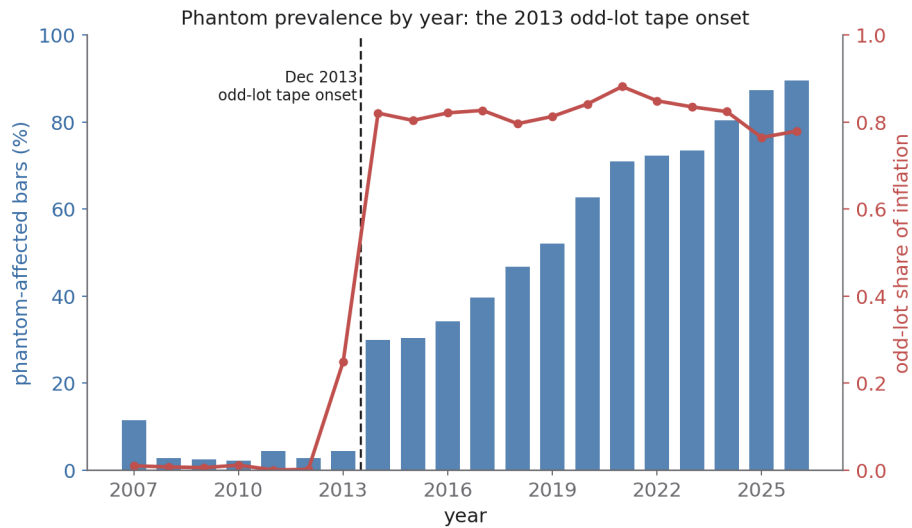
Year	Affected fraction	Median inflation (bp)	Odd-lot share
2007	0.115	10.6	0.01
2008	0.028	13.2	0.01
2010	0.021	6.3	0.01
2013	0.044	2.9	0.25
2014	0.300	1.9	0.82
2016	0.343	2.0	0.82
2018	0.468	2.4	0.80
2020	0.627	3.6	0.84
2022	0.723	4.0	0.85
2024	0.804	4.1	0.82
2026	0.896	7.0	0.78

Table 3: Phantom prevalence and magnitude by price band over the full universe. High-priced names carry the highest prevalence (odd-lot trading dominates when a round lot is expensive); low-priced names carry the largest relative magnitude (a fixed sub-penny displacement is a larger fraction of a low price). Population aggregates over affected bars. Source: `prevalence_fixed.json`.

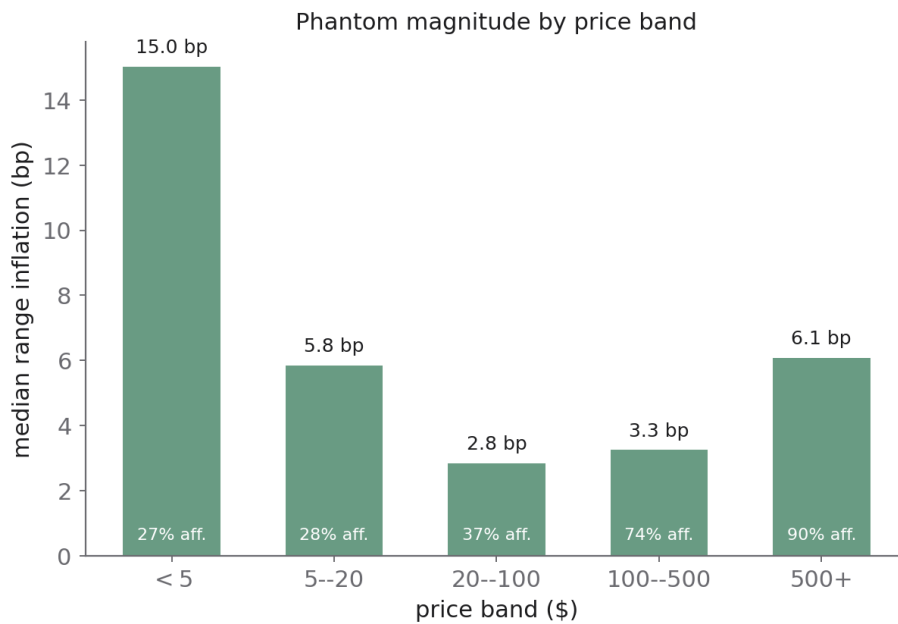
Price band	Bars (millions)	Affected fraction	Median inflation (bp)
< \$5	59	0.273	15.0
\$5–20	259	0.283	5.9
\$20–100	864	0.368	2.8
\$100–500	268	0.741	3.3
≥ \$500	15	0.901	6.1

by 2026, with odd lots supplying 78 to 85% of the range inflation throughout the odd-lot era (Table 2). The rise tracks the odd-lot share of the tape becoming a majority of prints by count over the same period. This by-year surface is itself a natural experiment: the phantom mechanism switches on with the odd-lot tape and grows with it, which is why the selection experiment is confined to 2014 onward.

The magnitude of the contamination, the naive range minus the executable range over affected bars, has a median of 3.4bp and a 99th percentile of 55.8bp; the mean exceeds the median because the contamination sits in a heavy tail. It is strongly price-dependent (Table 3). Low-priced names carry the largest relative inflation, a median of 5.9bp in the \$5–20 band and 15.0bp below \$5, because a fixed sub-penny displacement is a larger fraction of a low price. High-priced names carry the highest prevalence, 74% of bars affected in the \$100–500 band and 90% above \$500, because a 100-share round lot is prohibitively large there so odd-lot trading dominates. The two effects, larger relative magnitude at low prices and higher prevalence at high prices, are the two faces of the same odd-lot mechanism.



(a) Prevalence by year (the 2013 onset).



(b) Magnitude by price band.

Figure 2: Untradeable extremes over the full universe. Panel (a): the phantom-affected-bar share by year (bars) and the odd-lot share of the inflation (line), near zero before the December 2013 odd-lot tape onset and rising from 30% to 90% across the odd-lot era. Panel (b): the median range inflation by price band, largest in relative terms at low prices and highest in prevalence at high prices.

7 Results II: Decomposition of Range Inflation by Print Source

The decomposition here uses a diagnostic cascade that is distinct from the headline executable bar. The headline TRADE-NB bar keeps reachable inside-NBBO prints and drops only out-of-NBBO prints (Section 5.1). The decomposition here is a different, more aggressive object: to attribute the range inflation to print sources, it drops prints by source in a cascade, and at the off-exchange step it drops every Exchange-equals-FINRA “D” print, including reachable inside-NBBO “D” prints that the executable bar retains. The cascade therefore over-drops relative to the executable bar, and its “D” share is a cascade quantity, the off-exchange contribution to the cascade’s range reduction, not a count of prints the executable bar excludes. With that scope stated, the stacked attribution of the total excess-above-executable range inflation over the full universe is 76.3% odd-lot prints, 18.4% off-exchange “D” prints, 4.6% other ineligible conditions,¹ and 0.7% residual bid-ask bounce, with the bounce baseline anchored to the Corwin-Schultz and EDGE estimators (Table 4).

The scale of the full universe separates the sources cleanly. Over 1.46 billion bars the shares carry negligible sampling error, so the odd-lot and off-exchange contributions are firmly separated: odd lots lead at 76.3% against off-exchange 18.4%, a margin of +57.9 percentage points. Odd lots are unambiguously the dominant source of the range inflation, consistent with their being the dominant source of the prevalence itself (Section 6) and with the 2013 tape onset that switches the mechanism on (Section 10). A finer split of the “D” bucket into its reachable-inside-NBBO and out-of-NBBO halves needs per-print NBBO matching and is future work. Because the executable extreme is reconstructed as the union of the executable and remaining prints, the excess-above-executable inflation is non-negative on every bar; there is no negative residual.

This Section 7 result, the magnitude split of the bar-range inflation by source, is the source-specificity evidence for the headline: odd lots supply three-quarters of the inflation, the same channel that the 2013 tape onset switches on and that the by-year prevalence tracks. Sections 11 and 12 confirm the effect is specific to the untradeable-print channel through the CME futures control (no odd-lot channel, gap essentially zero) and the random-drop placebo.

8 Results III: Selection Fidelity, the Headline

This is the lead result, and it is a result about selection fidelity, not profit. We state the framing plainly before the numbers, because the numbers do not mean what a casual reading might expect.

The universe is edge-free by construction. The corpus is 7,232 strategies spanning twelve structural families (Section 5.4), each an equal-weight cross-symbol portfolio over the full 998-name universe, scored net of real costs, and net of real costs nothing in it is profitable out-of-sample: the best clean-out-of-sample Sharpe is -0.13 , the median

¹The cascade’s ineligible step drops prior-reference (“P”), out-of-sequence, and derivatively-priced prints, which the corrected per-field eligibility (Table 1, Appendix A) shows do update the official high and low. This bucket is therefore an upper bound on the genuinely high/low-ineligible contribution, pending a per-field recomputation of the cascade.

Table 4: Cascade decomposition over the full universe: excess-above-executable range-inflation shares by print source. The shares are a diagnostic-cascade quantity, not an executable-bar count (see text). Population aggregates over 1.46 billion bars; sampling error is negligible and the odd-lot-versus-off-exchange margin is firmly separated. Source: `prevalence_fixed.json`.

Source	Share
Odd-lot	0.763
Off-exchange “D”	0.184
Other ineligible	0.046
Residual bid-ask bounce	0.007
Odd-lot minus off-exchange margin	+0.579

Phantom-print source decomposition (full universe)

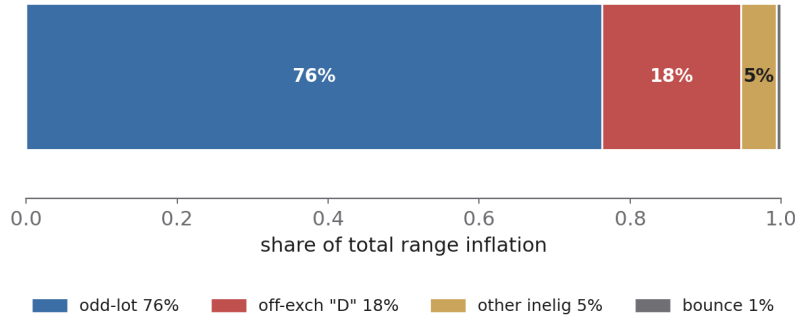


Figure 3: Stacked attribution of range inflation by print source over the full universe: odd-lot prints dominate at 76%, off-exchange “D” prints next at 18%, with other ineligible conditions and residual bid-ask bounce small. Over 1.46 billion bars the odd-lot-minus-off-exchange margin (+58 percentage points) is firmly separated.

−0.61, and not one strategy is positive (on the ≈ 20 mutually uncorrelated strategies the best is −0.30 and the median −0.60). This is by design. The cleanest place to measure what phantom-print contamination does is a universe where there is no real edge to find, because then any apparent in-sample edge is, by construction, false, and the only question is whether the backtest data lets that false edge propagate into selection. So the result below is not the discovery of a profitable strategy. It is the measurement of how much more faithfully a clean-bar backtest selects strategies than a naive-bar backtest, in a setting where the honest answer is that you should trade nothing.

Selection fidelity (the pre-registered primary endpoint). The endpoint is how faithfully an in-sample ranking of strategies predicts their executable out-of-sample ranking, measured by the in-sample-rank to clean-out-of-sample-rank information coefficient. Selecting on clean executable bars (Arm C) predicts the executable out-of-sample ranking with rank-IC 0.9987, and selecting on naive phantom-contaminated bars (Arm N) with 0.9948, a gap of +0.0039. Both arms are high because, in an edge-free universe, a strategy’s persistent cost-and-turnover signature dominates its Sharpe ranking in and out of sample; the quantity that matters is the naive-versus-clean gap, not the absolute level. Significance is the subtle part, because the 7,232 strategies are cross-symbol portfolios of the same specification families and so are highly correlated: an i.i.d. bootstrap over

Table 5: The headline gap under the effective- N block bootstrap versus the anticonservative i.i.d. diagnostic, plus the mutually uncorrelated subset. The block bootstrap resamples the ≈ 20 correlation blocks (each strategy assigned to its most-correlated de-duplicated representative), pricing significance at the effective number of independent bets; the i.i.d.-over-7,232 row is reported only to show how far it overstates precision. Same seed and 5,000 replications. Source: results.json.

Resampling unit	Units	Point est. arms	Gap	95% CI / p
Effective- N block bootstrap	≈ 20 blocks	0.9987 vs 0.9948	+0.0039	[0.0025, 0.0093], $p < 0.001$
i.i.d. over strategies (diagnostic)	7,232	0.9987 vs 0.9948	+0.0039	[0.0037, 0.0043], $p < 0.001$
Mutually uncorrelated subset	20	0.9925 vs 0.9729	+0.0196	$p = 0.06$

all 7,232 would treat them as independent bets and badly overstate precision. We instead assign each strategy to its most-correlated de-duplicated representative, forming ≈ 20 correlation blocks, and resample whole blocks (Table 5). The block bootstrap gives the gap a 95% interval of [0.0025, 0.0093] with $p < 0.001$; the i.i.d.-over-7,232 diagnostic gives the far tighter [0.0037, 0.0043], and the gap between the two interval widths is the factor by which treating correlated strategies as independent would overstate precision. On the 20 mutually uncorrelated strategies alone the gap is larger, +0.020 (0.9925 clean versus 0.9729 naive), but marginal at that sample size ($p = 0.06$); the all-strategy block estimate is the primary because it uses every strategy for the point estimate while pricing significance at the effective number of independent bets. Under both readings the gap is positive. The naive backtest is not noisy or broken: it ranks strategies, just less faithfully to what a live order would realize. Two cautions on reading the level. First, the absolute IC is high under both arms for the mechanical reason above: cost and turnover persist across the in-sample and out-of-sample windows, so the relative badness of strategies ranks stably even when nothing is profitable. Second, the gain is fidelity, not profit: a more faithful ranking of an edge-free universe still ranks losses.

Robustness across corpus construction. The fidelity gap is de-duplication-dependent in magnitude but never in sign. Sweeping the de-duplication ranking key and the correlation threshold τ , the rank-IC gap runs from +0.007 to +0.020, with the clean rank-IC always above the naive one (Table 6). The gap is significant at the looser thresholds that retain more strategies ($p \leq 0.006$ at $\tau = 0.7$, both keys) and marginal at $\tau = 0.5$ ($p \approx 0.06$ –0.15); the tightest $\tau = 0.3$ retains only six cross-symbol strategies, so mutually perfectly rank-correlated that both arms hit 1.000 and the gap is mechanically zero, which we report rather than suppress. Across every cell that retains more than a handful of strategies the gap is positive, and the all-strategy effective- N estimate of +0.0039 ($p < 0.001$) is the primary because it does not depend on a de-duplication threshold at all.

A signal-mechanism story does not survive testing. One might expect the contamination to concentrate in strategies that reach for bar extremes, reading a per-family gap pattern as a mechanism. The full universe does not support that reading: the per-family gaps are heterogeneous with wide, overlapping intervals and do not separate the families whose signals reference a bar extreme from those that reference the close or a moving average. The family-clustered gap over the nine of the twelve families that retain at least one strategy after de-duplication is +0.020 with a 95% interval of [0, 0.062] ($p = 0.12$), consistent with the headline but not attributable to any single signal family. The headline

Table 6: Selection-fidelity robustness sweep on the full universe. Each row is one de-duplication construction (ranking key and correlation threshold τ), with the number of mutually uncorrelated cross-symbol strategies, the naive and clean rank-IC, their gap, and the gap p -value under an i.i.d. bootstrap over the deduped set. The gap is positive in every cell; the tightest $\tau = 0.3$ retains only six strategies with both arms at 1.000 (mechanically zero gap) and is omitted. The all-strategy effective- N estimate ($+0.0039$, $p < 0.001$) does not depend on τ . Source: `results.json`.

Dedup key	τ	N	Naive IC	Clean IC	IC gap	p (gap)
mean-in-sample	0.5	20	0.9729	0.9925	+0.0196	0.061
mean-in-sample	0.7	70	0.9909	0.9979	+0.0070	< 0.001
trade-count	0.5	21	0.9792	0.9909	+0.0117	0.150
trade-count	0.7	69	0.9871	0.9958	+0.0086	0.006

gap stands on its own; it does not depend on a signal-family partition.

Source specificity (within-equity). The fidelity effect is specific to the untradeable prints, not to having fewer prints. The source decomposition of Section 7 localizes it: odd-lot prints supply 76% of the range inflation, off-exchange “D” prints 18%, other ineligible conditions 5%, and residual bid-ask bounce under 1%, so the phantom mechanism is overwhelmingly the odd-lot channel, the same channel whose December 2013 tape onset drives the by-year prevalence jump of Section 6 (near-zero phantom prevalence before 2014, rising to 90% by 2026). A random-print-drop placebo, removing the same count of prints per bar but chosen without regard to type, makes selection no better than naive (Section 12), ruling out a fewer-prints explanation, and the cross-market CME futures control (Section 11), which has no odd-lot or off-exchange channel, returns a fidelity gap of essentially zero. The effect tracks the untradeable-print channel specifically, which is the within-equity specificity evidence a cross-market control otherwise supplies.

The internal-validity control. On the 20 mutually uncorrelated strategies, the Arm N-home rank-IC, naive in-sample scored against its own naive out-of-sample, is 0.991, higher than the same naive selection scored against the executable out-of-sample (0.973). The naive backtest predicts its own contaminated out-of-sample better than the executable one. This is the crux: the naive backtest is internally consistent and externally less valid, because the phantom prints contaminate its in-sample and its naive out-of-sample identically, so it ranks faithfully against itself and less faithfully against the prices a live order could actually reach.

Deflation and overfitting. The deflated Sharpe ratio is zero for both arms: no strategy survives deflation by the number of trials, which is consistent with there being no real edge anywhere in the corpus and is the correct null answer for an edge-free universe. PBO via CSCV is not defined on the ≈ 20 -strategy uncorrelated set, which has too few configurations for the combinatorially-symmetric cross-validation to split; we do not report it, and note that the null deflated Sharpe already delivers the edge-free conclusion at the assembly step.

Selection value. A practitioner-facing question is how much realized performance the naive selection costs: the realized clean-out-of-sample Sharpe of the top decile by in-

sample Sharpe, Arm C minus Arm N. On the full universe this gap is essentially zero: both arms' in-sample top decile realizes an identical clean-out-of-sample Sharpe of -0.35 . With only ≈ 20 mutually uncorrelated strategies and no edge anywhere, the two arms select the same handful of least-bad strategies, so the realized-dollar gap is null even though the rank-IC gap is positive. This is the expected outcome in an edge-free universe: the fidelity gap is a ranking-corruption result, and the realized-dollar penalty of that corruption is nonzero only where the corruption changes which strategy you would actually deploy, which requires a universe with a deployable strategy to change to. The robust commercial claim, that clean bars let a genuinely profitable strategy survive while naive bars degrade it, therefore has no naturally occurring example in this edge-free universe. We establish it in two complementary ways: Theorem 3 supplies it as a proved consequence and Remark 3 evaluates the implied penalty at the measured contamination parameters, and a labelled known-answer planted-drift experiment (Section 13.8) finds the clean-selected book higher at every tested drift strength, consistent with a deployment flip it suggests rather than directly observes.

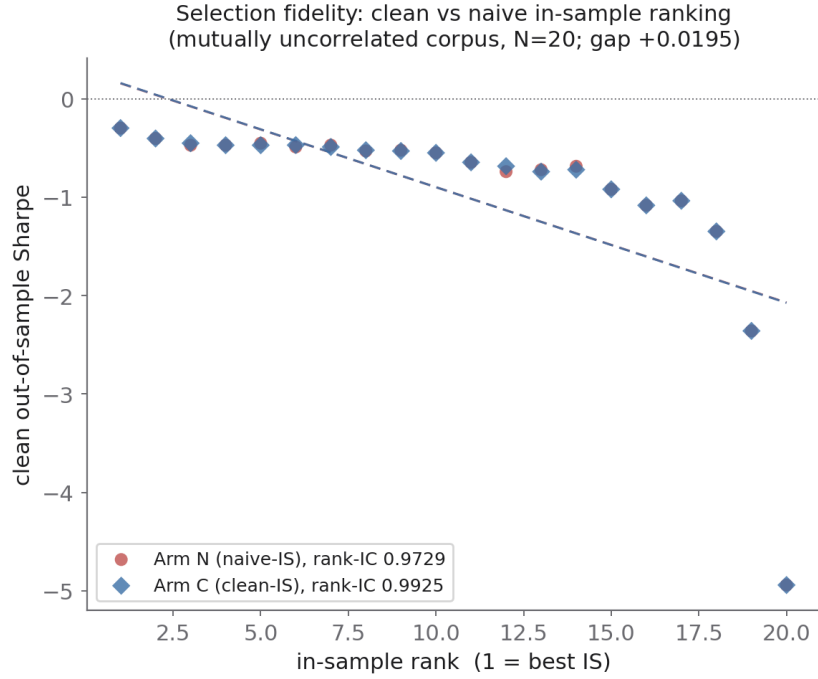
The replication ratio, demoted and explained. A naive in-sample-to-out-of-sample replication ratio (clean out-of-sample Sharpe over in-sample Sharpe) reads, at the median, 1.11 under naive selection and 1.07 under clean selection, a difference that points nowhere useful. We do not headline it, and the reason is that the ratio is confounded in this setting. When the in-sample Sharpe is itself the contaminated quantity, and when both numerator and denominator can be negative in an edge-free universe, the ratio is ill-defined as a fidelity measure: a strategy that looks badly negative in-sample and badly negative out-of-sample can post a high ratio that means nothing. The rank-IC and the selection-value metrics, which compare a fixed in-sample ranking or selection against the executable out-of-sample, are the correct measures here. We report the ratio transparently to avoid the appearance of suppressing an inconvenient number, but it is not evidence for or against the claim.

A footnote on survival. The survival fraction (of in-sample-profitable strategies, the share still profitable on the clean out-of-sample) is undefined under the headline corpus, because no in-sample-profitable strategy survives de-duplication in this edge-free universe; it takes a value only in some non-default cells. We therefore do not report it in the headline, consistent with the edge-free universe.

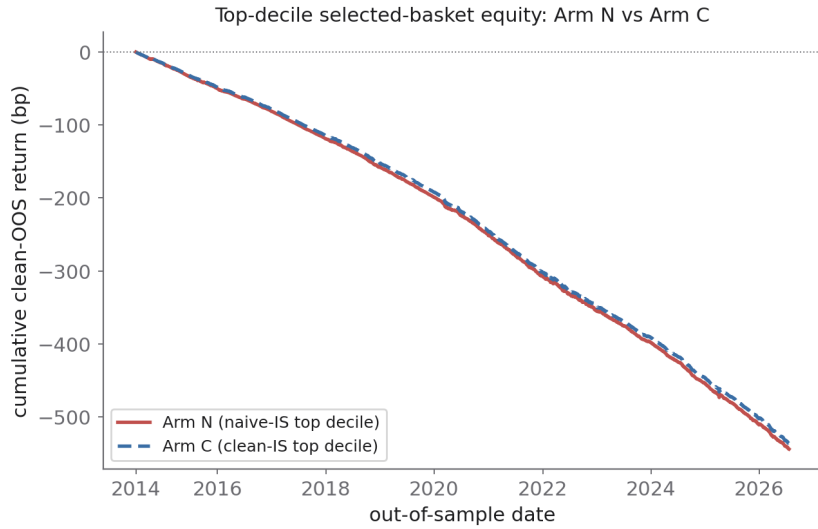
9 Results IV: Intrabar Ambiguity, Tick-Resolved

This section separates the bar-extreme contamination from the intrabar path assumption, and it confirms the headline is not an artifact of how the engine breaks intrabar ties. The measurement is deliberately clean: on nine symbol-days (NVDA, SPY, F across three 2023 sessions), both the one-minute bars and the tick path are built from the same executable TRADE-NB stream, so there are no phantom extremes in either, and what is left is pure intrabar-ordering ambiguity. For each triggered trade we compare three conventions: optimistic (the target is hit first on every ambiguous bar), pessimistic (the stop is hit first), and the tick-resolved truth (the executable ticks walked in time order, first level crossed). Costs are charged round-trip through the realism cost model.

Three findings. First, the ambiguity is rare and bracket-width-dependent: 3.15% of triggered trades pass through a bar where both the stop and the target lie inside the range,

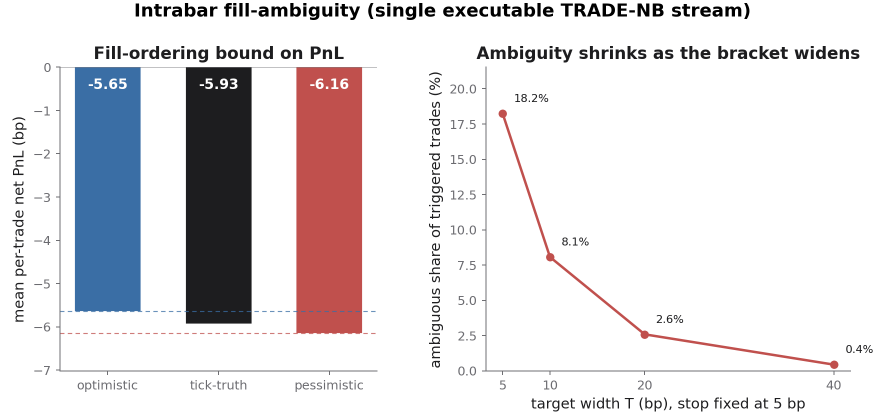


(a) In-sample rank to clean-OOS rank, both arms.



(b) Top-decile realized clean-OOS equity.

Figure 4: The headline, a selection-fidelity result in an edge-free universe. Panel (a) (the money figure): the in-sample ranking under clean selection (Arm C) tracks the executable out-of-sample ranking (rank-IC 0.9987), slightly better than naive selection (Arm N, 0.9948); the gap is small in level because a persistent cost signature dominates the ranking of an edge-free universe, and the decisive quantity is the naive-versus-clean gap. Panel (b): the realized clean-out-of-sample equity of the in-sample top decile under each arm; both lose identically in this edge-free universe, so the fidelity gap is a ranking-corruption result, not a realized-dollar one (Theorem 3 carries the dollar claim).



Per-trade PnL bounds and ambiguity vs bracket width; overall 3.2% of triggered trades ambiguous (nine symbol-days, tick-resolved).

Figure 5: Intrabar ambiguity, tick-resolved. Left: the mean per-trade net PnL under the optimistic, tick-truth, and pessimistic fill-ordering conventions; the tick-resolved truth (the centre bar) sits at 0.44 of the pessimistic-to-optimistic span, near the midway point and if anything leaning pessimistic, so optimistic tie-breaking overstates performance by under 1 bp per trade. Right: the ambiguous-trade share falls steeply as the stop/target bracket widens.

rising to 18.2% at tight 5 bp by 5 bp brackets and falling to at most 0.5% once either leg is 20bp or wider. Second, the convention gap is small: the optimistic-minus-pessimistic difference is 0.51 bp per trade (symbol-day-clustered bootstrap CI [0.232, 0.825], $p \approx 0$; a Sharpe gap of 0.024). Third, and most important, the tick-resolved truth sits at 0.444 of the span from the pessimistic to the optimistic extreme (CI [0.334, 0.555], which brackets 0.5 and confirms the truth is near the midway point, if anything leaning pessimistic): the pooled mean per-trade PnL is -5.65 bp optimistic, -5.93 bp at the tick truth, and -6.16 bp pessimistic. Of the ambiguous trades, 43.9% have a tick-resolved PnL that beats the pessimistic convention, so the real ordering is roughly a coin flip leaning slightly pessimistic, not the favorable-exit-first that an optimistic engine assumes.

The implication supports the headline directly. On bars with no phantom extremes, the most an optimistic convention can overstate performance is under 1 bp per trade, and the truth is close to pessimistic, so an optimistic-tie-breaking objection cannot explain the contamination the paper measures. The dominant backtest-realism effect is the bar-extreme phantom channel (median range inflation 3.4 bp, p99 55.8 bp, Sections 6 through 8), which is a separate and larger channel than the intrabar one measured here.

10 Robustness

This section stress-tests the headline beyond the corpus-construction sweep already reported in Section 8 (four dedup-key by τ cells, the fidelity gap positive in all). The cost model (a modeled time-of-day half-spread plus a per-fill commission, charged round-trip in both arms) is applied throughout, so every gap reported is already net of cost; because the naive bar trades at parity with the clean bar, the two arms carry near-identical cost, and the gap is a pure ranking-fidelity difference, not a cost artifact. The intrabar-convention robustness is in hand: the headline rank-IC fidelity gap is not an artifact of how intrabar ties are broken, because the selection experiment realizes its out-of-sample on executable bars under a single consistent fill model in both arms, and the intrabar channel measured in

Table 7: Cost-multiplier stress on the full-universe headline. The per-fill friction (time-of-day half-spread plus commission, charged round-trip in both arms) is scaled by the multiplier and the all-strategy rank-IC gap (Arm C minus Arm N) is recomputed, with the 95% interval from the effective- N block bootstrap (≈ 20 correlation blocks, 5,000 replications, same seed). The gap stays positive and significant through $5\times$. Source: `results.json`.

Cost multiplier	Rank-IC gap	95% CI	p
$1\times$ (modeled)	+0.0039	[0.0025, 0.0093]	< 0.001
$2\times$	+0.0029	[0.0017, 0.0077]	< 0.001
$3\times$	+0.0031	[0.0015, 0.0085]	< 0.001
$5\times$	+0.0031	[0.0013, 0.0108]	< 0.001

Section 9 is small (an optimistic-minus-pessimistic gap of 0.51 bp per trade) and lands near the midway point (tick truth at 0.444 of the span), so an optimistic-convention objection does not explain the gap.

Cost-multiplier stress. Scaling the per-fill friction (the modeled time-of-day half-spread plus commission) by $2\times$, $3\times$, and $5\times$ in both arms and recomputing the all-strategy rank-IC gap on the full universe leaves the gap positive and significant at every level (Table 7). The gap is near-invariant to the cost level: it reads +0.0039 at the modeled cost, +0.0029 at $2\times$, and +0.0031 at both $3\times$ and $5\times$, with every 95% effective- N block-bootstrap interval strictly above zero and $p < 0.001$ throughout. The invariance is the mechanism rather than a coincidence: because the naive and clean arms trade at parity, a cost multiplier debits both arms near-identically, so it moves the level of each arm’s rank-IC but not the fidelity difference between them, and the gap does not approach zero even at five times the modeled cost.

Sub-period stability. Splitting the executable out-of-sample by calendar year and re-computing the all-strategy rank-IC gap in each, the gap is positive in every one of the thirteen years of the odd-lot era, ranging from +0.0002 in the partial 2026 to +0.0095 in 2015, with a typical value near +0.005. These are single-year out-of-sample estimates, a noisier and differently pooled quantity than the full-period +0.0039 headline rather than a re-estimate of it: single-year out-of-sample Sharpes are noisier, so the per-year rank-ICs (roughly 0.88 to 0.95) sit below the full-period figure, but the naive-versus-clean ordering never reverses, and clean selection is the more faithful arm in all thirteen years.

Survivorship. The universe is point-in-time and includes delisted names: a strategy is selected and scored only on the names tradeable on each date, so survivorship bias is eliminated by construction rather than tested as an add-on. This matters for the direction of the effect. Failed and distressed tapes are the most phantom-contaminated, because a name in terminal decline trades thin and volatile so a larger share of its bars carry a phantom extreme, and the naive arm degrades most on exactly those tapes. Because those names are already in the corpus on the dates they traded, the point-in-time construction is conservative for the fidelity gap, not optimistic: excluding them, as a survivor-only panel would, would understate how much cleaning helps.

Natural experiment: the 2013 odd-lot tape onset. This is the cleanest causal result, visible directly in the by-year prevalence of Section 6. Before December 9, 2013, odd-lot executions were not disseminated to the consolidated tape (SEC Release 34-70794); the day they were, phantom contamination switched on. Across the full universe the phantom-affected-bar share is 2 to 12% in 2007–2012 with an odd-lot share of the inflation near zero, jumps to 30% in 2014 with the odd-lot share at 0.82, and climbs to 90% by 2026 (Table 2). This is a whole-universe regime change: the phantom mechanism is created by the odd-lot tape and grows with the odd-lot share of prints, which is direct causal confirmation that odd-lot prints drive the contamination.

Natural experiment: the round-lot redefinition. Around the Market Data Infrastructure round-lot redefinition (Release 34-90610, accelerated by 34-101070), the phantom-affected-bar share is flat, moving -1.8 percentage points, within noise. This is expected, not a surprising null: those rules changed quote dissemination and added a SIP best-odd-lot identifier, not the trade-tape’s fixed quantity-under-100 odd-lot test that our bars use, so no trade-tape prevalence shift was mechanically expected, and the round-lot compliance date is around November 2025 and soft. We report it flat rather than force a result. Because our bars use a fixed under-100-share test rather than the price-tiered round-lot definition that took effect in November 2025 (Rule 600(b): 40 shares above \$250, 10 above \$1,000, 1 above \$10,000), that test labels post-2025 tier-round-lots in the highest price bands as odd lots, so the 2026 odd-lot *attribution* share is an upper bound; the 2026 phantom *prevalence* is unaffected, since prevalence is the naive-versus-executable NBBO comparison and does not depend on print size.

11 Negative Control: CME Futures

The cross-market falsification asks whether the contamination is specific to equity microstructure. We run the identical twelve-family strategy universe through the identical two-arm pipeline on CME futures, five front-month roots (ES, NQ, CL, GC, ZN) over 1,250 contract-days in 2023. Futures have no odd lots and no off-exchange or internalized venue, but they are not free of away-from-market prints: the CME tape carries block trades, exchange-for-related-position (EFRP) trades, trade-at-settlement (TAS), and implied or spread-leg prints, all reported as non-aggressor trades with no lit-book aggressor side. These are the futures analog of the equity phantom print, a reported transaction a marketable order on the central book could not have reached. The clean filter keeps only aggressor-typed trades, which removes them. So this is a genuine two-arm test of the same kind of cleaning, not a degenerate one. Same strategies, same engine: if the equity gap were a pipeline artifact, it would reappear here.

It does not. On the futures corpus ($N = 441$ uncorrelated strategies), the rank-IC is 0.9771 under naive selection and 0.9770 under clean, a gap of -0.00011 with a bootstrap confidence interval of $[-0.00077, +0.0000056]$ that spans zero ($p = 0.076$, not significant). The futures gap is statistically indistinguishable from zero, against a positive and significant equity gap ($+0.0039$ over all strategies with $p < 0.001$, $+0.020$ on the uncorrelated subset). The mechanism is the prevalence: across the five-root, full-2023 corpus (486,326 regular-trading-hours bars) only 0.80% of futures bars are phantom-affected, that is, have a naive high or low that differs from the clean one, so 99.2% of futures bars are byte-identical naive versus clean (against 42.3% phantom-affected bars in equities,

Table 8: The two-arm cross-market control, equities versus CME futures. The comparison is of the naive-versus-clean gap within each market, not of the corpus sizes: the equity strategies are cross-symbol portfolios of the same specifications and so are far more mutually correlated, de-duplicating to fewer uncorrelated strategies (20) than the futures corpus (441). The futures phantom prints removed by cleaning are away-market non-aggressor prints (block, EFRP, TAS, implied or spread-leg), not session markers, on the phantom-heavy roots. Sources: `results.json`, `results_futures.json`, `per_root_phantom.json`, `futures_print_types.json`.

Market	<i>N</i> uncorr.	Naive IC	Clean IC	Fidelity gap	<i>p</i>	Phantom bars
US equities, all strategies (eff- <i>N</i> boot)	7,232 (raw)	0.9948	0.9987	+0.0039	< 0.001	42.3%
US equities, uncorrelated subset	20	0.9729	0.9925	+0.0196	0.06	42.3%
CME futures (5 roots, 2023)	441	0.9771	0.9770	−0.00011	0.076	0.80%

Table 8, about 53 times less). On the phantom-heavy roots the differing bars are driven by these away-market prints, not by session markers: on a representative-day sample, away-market non-aggressor prints with positive quantity are 99.6% of the dropped prints on CL and 98% on GC, and they set the naive extreme on 100% of those roots’ phantom bars. ES is the natural control, a single deep book with essentially no phantom (0.000%), where the few dropped prints genuinely are zero-quantity session markers. The phantom share varies by root from ES at 0.000% and NQ at 0.012% up to GC at 1.06% and CL at 2.81% on the commodity tapes, but even the worst root is far too rare, against the deep aggressor flow, to move a strategy ranking. Two honest notes. First, the Type and Flags fields do not sub-separate block from EFRP from TAS, so we report these together as away-market non-aggressor prints rather than claim a finer split. Second, the equity definition of an odd lot, a fill under 100 shares, does not transfer to futures, where nearly every fill is under 100 contracts, so that analog was deliberately excluded from the futures cleaning. This strengthens the control rather than weakens it: futures do have an away-market-print analog, the cleaning removes it, and the fidelity gap is still essentially zero, because these prints are rare relative to the deep aggressor flow. The contamination is equity-microstructure-specific and not a pipeline artifact, established by a genuine two-arm cross-market control rather than by the dose-response of Section 8 alone.

12 Corollary Nulls

This section reports the placebo that rules out the “just fewer prints” explanation, complementing the source decomposition of Section 7 and the futures negative control of Section 11. The placebo drops a random equal-count subset of prints rather than the untradeable subset, then runs the same three-arm rank-IC comparison against the executable out-of-sample. If the fidelity gain came merely from having fewer prints, the random drop would behave like the real cleaning. It does not: the placebo rank-IC (0.558) falls below both the naive (0.913) and the clean (0.998) arms (Table 9), because removing random prints adds noise to the bar extremes rather than removing the untradeable ones that set them. The random same-count drop is thus worse than naive, not better, so the “just fewer prints” explanation is ruled out: the fidelity gain is specific to the *identity* of the removed prints, not their count. This placebo is reported on the eight-name development panel; across the full universe the same specificity is carried by the source decomposition, which localizes three-quarters of the inflation to odd lots (Section 7), and by the CME futures

Table 9: Placebo three-arm comparison (random equal-count print drop), eight-name development panel. All arms select in-sample and are scored on the same executable clean out-of-sample. Source: placebo.json.

Selection arm	Rank-IC to executable OOS
Naive (all prints)	0.913
Placebo (random equal-count drop)	0.558
Clean (untradeable prints removed)	0.998

control, whose fidelity gap is essentially zero (Section 11).

13 A Measurement-Error Model of Phantom-Print Overfitting

The preceding empirical analysis compares rankings constructed from naive and executable bars. This section explains how contamination can impair selection for executable deployment and how a component that recurs across windows can make same-tape validation more reassuring than executable validation. These are distinct mechanisms: a deployment-fidelity gap does not, by itself, identify persistence. Strategy-dependent volatility scaling and window-specific score noise can also alter rankings.

We separate exact measurement identities, conditional population results, and empirical interpretation. The covariance results require finite second moments but not Gaussianity. A distribution-free selection result states the additional condition under which the clean ranking is preferable; a Gaussian benchmark supplies a closed-form magnitude. Neither benchmark is treated as an identified structural description of the empirical corpus. The analysis applies classical measurement-error and information-loss arguments to the two-tape setting [22, 9]; the Gaussian selection calculation is an application of normal regression and truncation [23].

13.1 Primitives and the executable-edge null

Fix a finite horizon $[0, T]$ and a filtered probability space satisfying the usual conditions. Let S_t denote the efficient *discounted price*, not its logarithm. All gains in the following lemma are discounted price gains, net of the specified execution costs.

Assumption 1 (Executable-edge null). *Under $\mathcal{H}_0$, S is a square-integrable martingale. An admissible strategy has bounded predictable holdings H , finitely many trades, and an integrable cumulative execution-cost process C that is adapted and nondecreasing. Its net gain is*

$$G_t = \int_0^t H_u dS_u - C_t.$$

The cost restriction applies to the strategy's actual or modelled fills, not merely to observed prints. A sufficient implementation condition is that a signed quantity q_i executed at price F_i at time τ_i satisfies $q_i(F_i - S_{\tau_i}) \geq 0$, with nonnegative fees and other charges. Here $q_i > 0$ denotes a purchase and $q_i < 0$ a sale.

Signals based on a completed bar can determine subsequent holdings, but cannot be used to establish exposure before that information is available. The null is a sufficient theoretical restriction, not an empirical assertion that the tested market is a martingale.

Lemma 1 (No positive expected executable gain under the null). *Under Assumption 1, for deterministic bar boundaries $t_{k-1} < t_k$ and a fixed reference capital $V_0 > 0$, define*

$$r_k^E = \frac{G_{t_k} - G_{t_{k-1}}}{V_0}.$$

Then

$$\mathbb{E}[r_k^E \mid \mathcal{F}_{t_{k-1}}] \leq 0.$$

The result also applies after a strategy has been selected using information available before the deployment interval, provided the resulting holdings remain admissible.

Proof. The stochastic integral of bounded predictable holdings against a square-integrable martingale is a martingale. Its increment has conditional mean zero. The increment of C is nonnegative, so subtracting it and dividing by $V_0 > 0$ yields the inequality. Selecting an admissible rule using past information does not change this argument. $\square$

The lemma concerns conditional expected gains, not the sign of each realised return or the unbiasedness of an estimated Sharpe ratio. In particular, observing no positive executable OOS Sharpe in a tested universe does not establish the absence of positive expected returns in its population.

Assumption 2 (Aligned bar comparison). *For the range comparison only, executable and naive bars use the same price units and timestamp-assignment rule. In every compared bar k , the executable print set E_k is nonempty and the naive set is $J_k = E_k \cup P_k$, where P_k contains the additional prints. The corresponding extrema are H_k^E, L_k^E and H_k^N, L_k^N .*

Set inclusion immediately gives

$$H_k^N \geq H_k^E, \quad L_k^N \leq L_k^E, \quad \Delta_k := (H_k^N - H_k^E) + (L_k^E - L_k^N) \geq 0.$$

Strict inflation occurs whenever an additional print lies strictly outside the executable range. This is a property of the aligned construction. If reporting-time and execution-time assignments differ, set inclusion must be checked explicitly: a displaced print can expand the destination bar as well as alter the origin bar. No one-sided conservativeness claim is made for those cases.

Range inflation need not increase a strategy's measured performance. Altered extremes can change entries, exits, stops, targets, and subsequent positions in different directions. The price-level construction therefore motivates a strategy-specific measurement error, but does not determine its mean, variance, or persistence.

13.2 Strategies and the measurement channels

For a fixed strategy n , window j , and bar k , write

$$u_{n,j,k} = r_{n,j,k}^N - r_{n,j,k}^E.$$

This difference includes any change in the strategy's realised trade path caused by its data input. With K_j bars in the window, define the finite-horizon expectation

$$m_{n,j}^r := \frac{1}{K_j} \sum_{k=1}^{K_j} \mathbb{E}[u_{n,j,k}], \quad v_{n,j,k} := u_{n,j,k} - m_{n,j}^r.$$

Then $K_j^{-1} \sum_k \mathbb{E}[v_{n,j,k}] = 0$. The realised residuals need not sum to zero. Expectations hold the strategy rules fixed and average over the stipulated data-generating process. Persistence is an additional restriction on how the strategy-specific perturbation behaves across windows; it is not implied by this decomposition.

An exact identity separates numerator and denominator effects. Let $\bar{r}_{n,j}^E, \bar{r}_{n,j}^N$ be the mean returns used in the two scores, and let $s_{n,j}^E, s_{n,j}^N > 0$ be their volatility denominators. Use the same positive scale factor α_j for both arms, and define

$$\begin{aligned} a_{n,j} &= \alpha_j \frac{\bar{r}_{n,j}^E}{s_{n,j}^E}, & x_{n,j} &= \alpha_j \frac{\bar{r}_{n,j}^N}{s_{n,j}^N}, \\ \kappa_{n,j} &= \frac{s_{n,j}^E}{s_{n,j}^N}, & b_{n,j} &= \alpha_j \frac{\bar{r}_{n,j}^N - \bar{r}_{n,j}^E}{s_{n,j}^E}. \end{aligned}$$

Consequently,

$$x_{n,j} = \kappa_{n,j}(a_{n,j} + b_{n,j}), \quad b_{n,j} = x_{n,j}/\kappa_{n,j} - a_{n,j}. \quad (1)$$

This identity is algebraic; it does not use a delta-method approximation. The perturbation $b_{n,j}$ is measured in score units and is not interchangeable with the per-bar return bias $m_{n,j}^r$. It can contain both recurrent distortion and finite-window estimation error.

Remark 1 (What volatility scaling does and does not imply). *There is no general reason for $\kappa_{n,j} \leq 1$, because*

$$\text{Var}(r^N) = \text{Var}(r^E) + \text{Var}(u) + 2\text{Cov}(r^E, u).$$

The covariance term can have either sign. A common positive κ_j preserves the ranking within window j ; strategy-specific $\kappa_{n,j}$ need not. Thus denominator heterogeneity is itself a possible ranking channel. An observed split between inflated and deflated total Sharpe differences does not identify the signs of the recurrent numerator component separately.

Equation (1) applies to the particular return window and score definition used. An average of window Sharpes is not a pooled-return Sharpe, and generally has no single aggregate κ . Scores from overlapping windows must retain their induced dependence when aggregated. The factor α_j specifies score units; it is not automatically a valid annualisation factor in the presence of serial dependence [16].

13.3 The finite-moment factor representation

For one specified IS/OOS comparison, consider the reduced form

$$\begin{aligned} a_n^{\text{IS}} &= g_n + \zeta_{n,I}, & a_n^{\text{OOS}} &= g_n + \zeta_{n,O}, \\ b_{n,I} &= \bar{b}_I + m_n + e_{n,I}, & b_{n,O} &= \bar{b}_O + \rho m_n + e_{n,O}, \\ x_n &= \kappa_I(a_n^{\text{IS}} + b_{n,I}), & w_n &= \kappa_O(a_n^{\text{OOS}} + b_{n,O}). \end{aligned} \quad (2)$$

The constants $\bar{b}_I, \bar{b}_O$ allow nonzero average contamination. They do not affect correlations or within-window rankings. The constants $\kappa_I, \kappa_O > 0$ can differ between windows but are common across strategies for the raw-score formulas below. With heterogeneous measured ratios, the same formulas instead describe the normalised scores $x_n/\kappa_{n,I}$ and $w_n/\kappa_{n,O}$; those are not generally the actual naive rankings.

Assumption 3 (Finite-moment benchmark). *The factors have finite second moments, with*

$$\begin{aligned} \mathbb{E}[g] &= \mu, & \text{Var}(g) &= \gamma^2 > 0, & \mathbb{E}[m] &= 0, & \text{Var}(m) &= \tau^2 \geq 0, \\ \text{Cov}(g, m) &= c, & c^2 &\leq \gamma^2 \tau^2, & 0 &\leq \rho \leq 1. \end{aligned}$$

The four residuals $\zeta_I, \zeta_O, e_I, e_O$ are mean-zero, mutually uncorrelated, and uncorrelated with (g, m) , with variances $\sigma_I^2, \sigma_O^2, v_I, v_O$, respectively. These restrictions are assumptions, not consequences of using disjoint windows. All score variances used below are positive. No cross-strategy independence or Gaussianity is imposed for the covariance results.

Write

$$\begin{aligned} A_I &= \gamma^2 + \sigma_I^2, & A_O &= \gamma^2 + \sigma_O^2, \\ D_I &= A_I + \tau^2 + v_I + 2c, & D_O &= A_O + \rho^2 \tau^2 + v_O + 2\rho c. \end{aligned}$$

Here m is a component shared by the two score windows, e_I, e_O are window-specific perturbations, and ρ is an OOS loading, not an autocorrelation. Under the benchmark,

$$\text{Cov}(b_I, b_O) = \rho \tau^2, \quad \text{Corr}(b_I, b_O) = \frac{\rho \tau^2}{\sqrt{(\tau^2 + v_I)(\rho^2 \tau^2 + v_O)}}. \quad (3)$$

The correlation is defined when both perturbation variances are positive. In particular, if $v_I = v_O = 0$, $\tau^2 > 0$, and $\rho > 0$, that correlation is one, even when $\rho < 1$. A temporally correlated residual would add to the covariance in (3); recurring contamination is not automatically a permanent strategy effect.

Remark 2 (Reduced form, Gaussianity, and the null). *The representation models score variation rather than deriving its parameters from print prevalence or trigger geometry. The shared realisation of g and the residual orthogonality restrictions require empirical assessment. They are not guaranteed by deterministic strategy rules.*

The mean μ is unrestricted in the covariance analysis. A nondegenerate Gaussian g cannot be interpreted as an everywhere nonpositive true executable edge. A score-level null illustration can instead impose $g \leq 0$ almost surely and conditionally mean-zero residuals, without Gaussianity; connecting that score interpretation to actual returns requires the specified estimator model. Lemma 1 separately establishes the executable-gain restriction under its price-and-fill assumptions. Gaussian assumptions introduced later are not asserted simultaneously with an everywhere nonpositive Gaussian g .

13.4 The deployment-fidelity gap

Define Pearson deployment fidelities

$$\phi_C = \text{Corr}(a^{\text{IS}}, a^{\text{OOS}}), \quad \phi_N = \text{Corr}(x, a^{\text{OOS}}).$$

Theorem 1 (Deployment fidelity with correlated and transient contamination). *Under Assumption 3 and common positive within-window scalings,*

$$\phi_C = \frac{\gamma^2}{\sqrt{A_I A_O}}, \quad \phi_N = \frac{\gamma^2 + c}{\sqrt{D_I A_O}}. \quad (4)$$

If $c = 0$ and $\tau^2 + v_I > 0$, then $\phi_C > \phi_N$, independently of κ_I . If $\gamma^2 + c \geq 0$, the necessary and sufficient condition for strict superiority is

$$\gamma^4(\tau^2 + v_I) > 2\gamma^2(A_I - \gamma^2)c + A_I c^2. \quad (5)$$

If $\gamma^2 + c < 0$, then $\phi_N < 0 < \phi_C$.

Proof. The clean covariance is γ^2 . The naive-target covariance is $\kappa_I(\gamma^2 + c)$, and the naive variance is $\kappa_I^2 D_I$, giving (4). For nonnegative numerators, the sign of $\phi_C - \phi_N$ is the sign of

$$\gamma^4 D_I - A_I(\gamma^2 + c)^2 = \gamma^4(\tau^2 + v_I) - 2\gamma^2(A_I - \gamma^2)c - A_I c^2.$$

The stated cases follow. $\square$

Under orthogonal contamination, persistent and transient IS perturbations both lower deployment fidelity through their combined variance. The fidelity gap therefore does not separate those components. With $c \neq 0$, contamination can contain information about executable performance, and the clean-versus-naive ordering is conditional on (5), not universal.

Heterogeneous denominators. Even absent additive contamination, let $x = K a^{\text{IS}}$, where $K > 0$ is independent of the clean score pair, $\mathbb{E}[K] = \bar{\kappa}$, and $\text{Var}(K) = s_K^2$. Then

$$\begin{aligned} \text{Cov}(x, a^{\text{OOS}}) &= \bar{\kappa} \gamma^2, \\ \text{Var}(x) &= \bar{\kappa}^2 A_I + s_K^2 \mathbb{E}[(a^{\text{IS}})^2]. \end{aligned}$$

Inverting an orthogonal common- κ model would falsely attribute the additional score variance to

$$\tau_{\text{pseudo}}^2 = \frac{s_K^2}{\bar{\kappa}^2} \mathbb{E}[(a^{\text{IS}})^2],$$

although the example has no additive drift. A satisfactory single-slope regression does not eliminate this confounding. Without restrictions on the dependence of K on the other factors, the inferred drift is not even a general upper bound.

13.5 Same-tape fidelity and the deployment trap

Theorem 2 (When same-tape validation is more flattering). *Under the finite-moment benchmark,*

$$\phi_{\text{home}} := \text{Corr}(x, w) = \frac{\gamma^2 + (1 + \rho)c + \rho\tau^2}{\sqrt{D_I D_O}}. \quad (6)$$

For $c = 0$, same-tape fidelity exceeds deployment fidelity if and only if

$$\rho\tau^2(2A_O\gamma^2 + A_O\rho\tau^2 - \rho\gamma^4) > \gamma^4 v_O. \quad (7)$$

Consequently, if $c = 0$, $v_O = 0$, $\rho > 0$, and $\tau^2 > 0$, the inequality is strict even when $\sigma_I^2 \neq \sigma_O^2$. Additional naive-specific OOS noise can overturn it.

Proof. The covariance of the unscaled naive scores is $\gamma^2 + (1 + \rho)c + \rho\tau^2$, proving (6). For $c = 0$, both correlation numerators are positive. After squaring and cancelling D_I , the comparison is equivalent to

$$A_O(\gamma^2 + \rho\tau^2)^2 - \gamma^4(A_O + \rho^2\tau^2 + v_O) > 0,$$

which expands to (7). When $v_O = 0$, its left-hand side before subtracting $\gamma^4 v_O$ is strictly positive because $A_O \geq \gamma^2$ and $0 < \rho \leq 1$. $\square$

For $c \neq 0$, the sign is determined directly by (6) and (4); it need not follow the orthogonal case. When both relevant numerators are positive, it is equivalently determined by

$$A_O\{\gamma^2 + (1 + \rho)c + \rho\tau^2\}^2 > (\gamma^2 + c)^2 D_O.$$

Dependence between windows. If $\text{Cov}(\zeta_I, \zeta_O) = \omega \neq 0$ but the other exclusions are maintained, the three covariance numerators become $\gamma^2 + \omega$, $\gamma^2 + \omega + c$, and $\gamma^2 + \omega + (1 + \rho)c + \rho\tau^2$. The variances are unchanged. The orderings must then be re-evaluated; disjoint observations alone do not imply $\omega = 0$. Likewise, $\text{Cov}(e_I, e_O) \neq 0$ adds directly to the same-tape numerator. Serial-dependence-robust uncertainty estimates do not restore a misspecified population covariance restriction.

Corollary 1 (A null-compatible deployment trap). *Consider the additional score-level restrictions $g \leq 0$ almost surely, $\mathbb{E}|g| < \infty$, $g \perp m$, and $\rho > 0$. Suppose the two residual sums $U_I = \zeta_I + e_I$ and $U_O = \zeta_O + e_O$ are independent of each other and of (g, m) , with proper distributions, and that m has unbounded upper support. Let*

$$X = x/\kappa_I - \bar{b}_I = g + m + U_I, \quad W = w/\kappa_O - \bar{b}_O = g + \rho m + U_O.$$

For finite population upper- q thresholds t_I, t_O , define $B_q = \{X \geq t_I, W \geq t_O\}$. If

$$p_I(g, m) = \Pr(U_I \geq t_I - g - m), \quad p_O(g, m) = \Pr(U_O \geq t_O - g - \rho m),$$

then

$$\Pr(B_q) = \mathbb{E}[p_I p_O], \quad \mathbb{E}[g \mid B_q] = \frac{\mathbb{E}[g p_I p_O]}{\mathbb{E}[p_I p_O]} \leq 0 \quad (8)$$

whenever the denominator is positive. Moreover, along the upper support of m , $\Pr(B_q \mid m = M) \rightarrow 1$ as $M \rightarrow \infty$.

Proof. Conditional independence gives the product of the two survival probabilities. Integrating yields (8); its numerator is nonpositive because $g \leq 0$. For each finite g , both conditional thresholds for U_I, U_O tend to $-\infty$ as $M \rightarrow \infty$. Independence from m and dominated convergence give the limiting probability. $\square$

Equation (8) quantifies certification at a specified selection fraction without equating a high naive validation score with executable edge. For an additional untouched score $T^+ = g + \zeta^+$ satisfying $\mathbb{E}[\zeta^+ \mid g, m, U_I, U_O] = 0$, its expected value after certification is $\mathbb{E}[g \mid B_q] \leq 0$. This statement does not automatically apply to the contemporaneous a^{OOS} , whose noise also enters w . The null-compatible illustration is not the Gaussian cross-section used below, and no numerical trap probability is inferred without its factor distributions.

13.6 Selection value, genuine edges, and the Gaussian benchmark

A positive population-average signature is not necessary for some strategies to have positive expected performance. Nor do the fidelity results require a nonpositive population mean. At the price level, a genuine-edge alternative must be distinguished from a change in the mean of a score model.

Assumption 4 (Possible executable-edge alternative). *Beyond $\mathcal{H}_0$, permit $S = M + \Lambda$, where M is a square-integrable martingale and Λ is a predictable finite-variation process, retaining the admissible holdings and fill-cost restrictions. This permits, but does not guarantee, positive expected net executable gains. The resulting score distribution and its covariance structure are separate modelling questions; a profitable subset need not imply $\mu > 0$.*

Assumption 5 (Conditional-mean sufficiency for selection). *Let $S_C = a^{\text{IS}}$, $S_N = x$, and $T = a^{\text{OOS}}$, with T integrable. For the following distribution-free result, assume*

$$\mathbb{E}[T \mid S_C, S_N] = h(S_C),$$

where h is nondecreasing. Thus the naive score supplies no additional conditional-mean information about the target after observing the clean score, and larger clean scores imply no smaller conditional expected targets. Selection thresholds give each arm probability $q \in (0, 1)$; independent random tie-breaking can be used when needed.

Proposition 2 (Selection dominance without Gaussian tails). *Under Assumption 5, let C_q and N_q denote selection into the top- q groups by S_C and S_N . Then*

$$\mathbb{E}[T \mid C_q] - \mathbb{E}[T \mid N_q] \geq 0.$$

The inequality is strict if the selections disagree with positive probability where $h(S_C)$ differs from its threshold value.

Proof. Let t_C be the clean threshold. Conditional-mean sufficiency and $\Pr(C_q) = \Pr(N_q) = q$ give

$$\begin{aligned} q\{\mathbb{E}[T \mid C_q] - \mathbb{E}[T \mid N_q]\} &= \mathbb{E}[h(S_C)(\mathbf{1}_{C_q} - \mathbf{1}_{N_q})] \\ &= \mathbb{E}[(h(S_C) - h(t_C))(\mathbf{1}_{C_q} - \mathbf{1}_{N_q})]. \end{aligned}$$

Above the clean threshold both factors in the product are nonnegative; below it both are nonpositive. At the threshold the first factor is zero. The strictness condition follows from positivity of the product on a set of positive probability. $\square$

This is a conditional information-loss result, in the spirit of comparison-of-experiments arguments [9], not a consequence of a Pearson correlation ordering alone. The condition holds, for example, when S_N is generated from S_C by noise independent of (S_C, T) , and $\mathbb{E}[T \mid S_C]$ is nondecreasing. It need not hold when contamination carries incremental information about executable performance.

A useful non-Gaussian example is $S_C = g + \zeta_I$ and $T = g + \zeta_O$, with independent residuals independent of g , $\mathbb{E}[\zeta_O] = 0$, and $\zeta_I \sim \mathcal{N}(0, \sigma_I^2)$ for $\sigma_I^2 > 0$. If the naive score is generated from S_C using noise independent of (S_C, T) , then $h(s) = \mathbb{E}[g \mid S_C = s]$ and

$$h'(s) = \frac{\text{Var}(g \mid S_C = s)}{\sigma_I^2} \geq 0.$$

The identity follows by differentiating the Gaussian-likelihood posterior mean. Thus the selection result permits any finite-variance distribution of g , including one supported entirely on nonpositive values; it does not require a Gaussian upper tail for the strategy population.

The finite-universe analogue requires the corresponding *whole-panel* restriction $\mathbb{E}[T_n | \mathbf{S}_C, \mathbf{S}_N] = h(S_{C,n})$. Conditional on those panels, choosing the k largest clean scores maximises the sum of conditional expected targets among all size- k selections. Cross-strategy dependence is allowed under that restriction; marginal bivariate conditions alone do not establish it.

Lemma 2 (Gaussian selection value). *If (S, T) is jointly Gaussian with positive variances, let $z_q = \Phi^{-1}(1 - q)$ and $\lambda(q) = \phi(z_q)/q$, where Φ and ϕ are the standard normal CDF and density. Then*

$$\mathbb{E}[T | S \geq Q_{1-q}(S)] = \mathbb{E}[T] + \text{Corr}(S, T) \sqrt{\text{Var}(T)} \lambda(q).$$

Proof. With $Z = (S - \mathbb{E}[S])/\sqrt{\text{Var}(S)}$, Gaussian regression gives $\mathbb{E}[T | Z] = \mathbb{E}[T] + \text{Corr}(S, T) \sqrt{\text{Var}(T)} Z$. Integrating over $Z \geq z_q$ and using $\mathbb{E}[Z | Z \geq z_q] = \phi(z_q)/q$ proves the result. This is the one-threshold normal-truncation calculation [23]. $\square$

Theorem 3 (Average selected-score shortfall under Gaussian pairs). *Assume the finite-moment benchmark with common within-window scaling, and suppose the two pairs $(a^{\text{IS}}, a^{\text{OOS}})$ and (x, a^{OOS}) are jointly Gaussian. At population selection fraction q , their difference in expected selected executable OOS scores is*

$$\Delta_S(q) := \mathbb{E}[a^{\text{OOS}} | C_q] - \mathbb{E}[a^{\text{OOS}} | N_q] = (\phi_C - \phi_N) \sqrt{A_O} \lambda(q). \quad (9)$$

Under the finite-moment benchmark this equals

$$\lambda(q) \left\{ \frac{\gamma^2}{\sqrt{A_I}} - \frac{\gamma^2 + c}{\sqrt{D_I}} \right\}.$$

In the orthogonal common-scaling case $c = 0$, it is positive whenever $\tau^2 + v_I > 0$, increases strictly with that total IS perturbation variance, and is independent of μ , κ_I , and ρ . It increases as the Gaussian selection fraction q decreases.

Proof. Apply Lemma 2 to the two score-target pairs and subtract. Their common target mean cancels. For $c = 0$, $D_I = A_I + \tau^2 + v_I$, so the expression is $\gamma^2 \lambda(q) (A_I^{-1/2} - D_I^{-1/2})$. Its derivative with respect to the total perturbation variance is $\gamma^2 \lambda(q) / (2D_I^{3/2}) > 0$, and the upper-tail mean $\lambda(q)$ is strictly decreasing in q . $\square$

The target is an *average selected-strategy score*. The theorem is not a pathwise guarantee, a uniquely persistent-drift penalty, or a portfolio-Sharpe identity. For non-Gaussian pairs the magnitude must be obtained from the two conditional selection values, not by replacing $\lambda(q)$ with a single empirical clean-arm lift. A clean-versus-naive Pearson ordering alone does not determine the two upper-tail conditional means.

Finite selection sets. If the N candidate triples are independent and identically distributed with the stated Gaussian pair marginals, and each arm selects exactly k candidates, (9) holds with

$$\lambda_{N,k} := \frac{1}{k} \sum_{j=N-k+1}^N \mathbb{E}[Z_{(j)}]$$

instead of $\lambda(k/N)$, where $Z_{(j)}$ are the order statistics of N independent standard normals. For example, $\lambda_{20,2} = 1.637540$ and $\lambda_{20,1} = 1.867475$, whereas $\lambda(0.10) = 1.754983$ and $\lambda(0.05) = 2.062713$. This follows by conditioning each selected target on its own score and averaging its concomitant order statistics. An effective number of dependence blocks cannot automatically replace the actual selection universe in this formula.

Corollary 2 (Opposite signs of expected selected scores). *Let $L_C = \phi_C \sqrt{A_O} \lambda(q)$ and $L_N = \phi_N \sqrt{A_O} \lambda(q)$ in the Gaussian benchmark, with $L_C > L_N$. The clean-selected average has positive expectation and the naive-selected average has negative expectation precisely when*

$$-L_C < \mu < -L_N.$$

The interval has width $L_C - L_N = \Delta_S(q)$. In the orthogonal case both lifts are positive, so this interval has a negative population mean; it can still contain a profitable selected subset.

Proof. The two expected selected scores are $\mu + L_C$ and $\mu + L_N$. Solving their opposing strict sign inequalities gives the interval and its width. $\square$

Remark 3 (From strategy scores to portfolio economics). *An average of constituent Sharpes is not generally the Sharpe of their combined return stream. For fixed portfolio weights w , expected return vector μ_r , and return covariance matrix Σ , measured at the same horizon,*

$$\text{SR}_{\text{book}}(w) = \frac{w^\top \mu_r}{\sqrt{w^\top \Sigma w}}.$$

A valid dollar comparison requires executable portfolio return paths, specified aggregation and rebalancing rules, costs, and a consistent annualisation convention [16]. Only when $\Delta \text{SR}_{\text{book}}$ is a difference in true annual book Sharpes at a common true annual volatility σ_{tgt} does $K \sigma_{\text{tgt}} \Delta \text{SR}_{\text{book}}$ give the corresponding annual expected excess-P&L difference on capital K . No such portfolio-level amount is identified from (9) alone. Parameterised rules can be included in the candidate universe; additional adaptive retuning outside that predefined selection problem is not covered by the theorem.

13.7 Dispersion, identification, and empirical scope

Proposition 3 (Dispersion is not an expected-maximum ordering). *Under the common-scaling benchmark,*

$$\sigma_{N,I} = \kappa_I \sqrt{D_I}, \quad \sigma_{E,I} = \sqrt{A_I}, \quad \sigma_{N,O} = \kappa_O \sqrt{D_O}.$$

Thus IS compression is equivalent to $\kappa_I < \sqrt{A_I/D_I}$. It does not, by itself, establish an ordering of the uncentred expected maxima.

Proof. The dispersion expressions follow from the model variances. To see why compression alone is insufficient, under independent Gaussian candidate scores let $b_N = \mathbb{E}[\max_{1 \leq n \leq N} Z_n]$. Then

$$\begin{aligned}\mathbb{E}[\max_n a_n^{\text{IS}}] &= \mu + \sqrt{A_I} b_N, \\ \mathbb{E}[\max_n x_n] &= \kappa_I(\mu + \bar{b}_I) + \kappa_I \sqrt{D_I} b_N.\end{aligned}$$

Their difference includes the mean term $(\kappa_I - 1)\mu + \kappa_I \bar{b}_I$ as well as the dispersion term. $\square$

The exact independent-normal expectation is

$$b_N = N \int_{-\infty}^{\infty} z \varphi(z) \Phi(z)^{N-1} dz.$$

For $N = 20$, $b_N = 1.867475$, not $\sqrt{2 \log 20} = 2.447747$. The latter is an upper bound and leading asymptotic scale, not a reliable small- N point approximation. Dispersions 0.480 and 0.438 would give centred Gaussian expected maxima 0.896388 and 0.817954 for twenty independent candidates, before adding the appropriate means. These are conditional illustrations, not measured maxima. Different tails and dependence require a joint model or resampling of the actual candidate panel. Compression therefore does not establish the absence of a multiple-testing problem [5].

What paired scores identify. With the underlying return means and denominators, (1) measures each $\kappa_{n,j}$ and $b_{n,j}$ directly. Under the factor exclusions, the paired-score covariances additionally satisfy

$$\text{Cov}(a^{\text{IS}}, b_I) = \text{Cov}(a^{\text{OOS}}, b_I) = c, \quad \text{Cov}(a^{\text{IS}}, b_O) = \rho c.$$

These are model restrictions to test, not consequences of residualising a regression. A fitted ordinary least-squares residual is orthogonal to its regressor by construction.

For a constant κ_I , the population regression of $x - a^{\text{IS}}$ on a^{IS} , with an intercept, has

$$\beta = \kappa_I - 1 + \frac{\kappa_I c}{A_I}, \quad \text{Var}(\text{residual}) = \kappa_I^2 \left(\tau^2 + v_I - \frac{c^2}{A_I} \right).$$

Thus its slope is not generally a pure attenuation estimate, and its residual variance does not generally identify persistent drift.

Two-window identification requires additional restrictions. Even in the orthogonal submodel $c = 0$, the two normalised perturbations supply

$$V_I := \text{Var}(b_I) = \tau^2 + v_I, \quad V_O := \text{Var}(b_O) = \rho^2 \tau^2 + v_O, \quad P := \text{Cov}(b_I, b_O) = \rho \tau^2.$$

These are three moments for four unknowns. If $P > 0$, the identified set under $0 < \rho \leq 1$ and nonnegative residual variances is

$$\max \left(P, \frac{P^2}{V_O} \right) \leq \tau^2 \leq V_I, \quad \rho = \frac{P}{\tau^2}, \quad v_I = V_I - \tau^2, \quad v_O = V_O - \frac{P^2}{\tau^2}, \quad (10)$$

provided the interval is nonempty. The lower bounds enforce $\rho \leq 1$ and $v_O \geq 0$, and the upper bound enforces $v_I \geq 0$. At $P = 0$, absence of a shared covariance does not identify τ^2 unless the loading is restricted. When $c \neq 0$, the additional cross-moment ratio $\text{Cov}(a^{\text{IS}}, b_O)/\text{Cov}(a^{\text{IS}}, b_I) = \rho$ can identify the loading under the same exclusions, but is unstable near $c = 0$. Correlated residuals or random scaling require further modifications.

Additional windows can identify a more restrictive repeated-measurement model. For example, if centred perturbations obey $b_j = \ell_j m + e_j$ for three windows, $\ell_1 = 1$, and all residuals are mutually uncorrelated and uncorrelated with m , then

$$\tau^2 = \frac{\text{Cov}(b_1, b_2) \text{Cov}(b_1, b_3)}{\text{Cov}(b_2, b_3)},$$

when the denominator is nonzero. The formula is conditional on the one-factor and residual exclusions; overlapping windows or additional shared factors can invalidate it.

A restricted calibration, not a structural finding. Only under common scaling, $c = 0$, no naive-specific residuals, uncorrelated clean residuals, and equal clean IS/OOS variances does the original inversion reduce to

$$\frac{\gamma^2}{A} = \phi_C, \quad \frac{\tau^2}{A} = \left(\frac{\phi_C}{\phi_N} \right)^2 - 1, \quad \kappa_I = \frac{\sigma_{N,I} \phi_N}{\sigma_{E,I} \phi_C}.$$

The same-tape correlation then determines ρ within that restricted family, when an admissible solution exists. A marginal variance difference alone remains underdetermined, and allowing v_I, v_O or denominator heterogeneity destroys the claimed unique persistence interpretation.

The reported rank correlations for the approximately twenty-strategy uncorrelated subset are $r_{s,C} = 0.9925$, $r_{s,N} = 0.9729$, and $r_{s,\text{home}} = 0.9910$. The reported primary all-strategy rank comparison, 0.9987 versus 0.9948, is a different panel and must not be combined with subset moments as though they came from one covariance matrix. Pairwise uncorrelated returns also do not establish independence of estimated strategy scores.

For bivariate normal scores, and only for that benchmark when interpreting raw-score Pearson correlations,

$$r_s = \frac{6}{\pi} \arcsin(\phi/2), \quad \phi = 2 \sin(\pi r_s/6).$$

Converting the subset values gives $\phi_C = 0.993191$, $\phi_N = 0.975323$, and $\phi_{\text{home}} = 0.991827$. Without bivariate normality, Pearson correlations should be computed directly from the paired scores. A Gaussian copula alone does not make the transformed parameter the Pearson correlation of arbitrary raw-score marginals.

For illustration, additionally suppose the clean IS and OOS standard deviations are both 0.480 and the naive IS standard deviation is 0.438, all on that same panel and in common units. Normalising $\gamma^2 = 1$, the restricted benchmark yields

$$\sigma_I^2 = \sigma_O^2 = 0.006856, \quad \tau^2 = 0.037228, \quad \rho = 0.702913, \quad \kappa_I = 0.896084.$$

Here $\tau^2/A = 0.036975$ is a *benchmark-implied* ratio, not an established share of persistent phantom variance. The implied population shortfalls are 0.015051 at $q = 0.10$, 0.017691 at $q = 0.05$, and 0.022858 at $q = 0.01$. Selecting two of twenty independent Gaussian candidates instead gives 0.014044. These are conditional score-unit calculations, not confidence intervals, bounds, or dollar losses.

Empirical requirements and uncertainty. The score-tail shape, the conditional-mean restriction, denominator heterogeneity, and the residual exclusions must be assessed using the paired strategy-by-window observations. Gaussian calibration should be checked against the observed conditional lifts of *both* arms, not only their rank correlations. Under the illustrative calibration, the clean top-decile lift above its population mean is 0.836656; that is a prediction to test, not a fitted empirical fact. A median is not interchangeable with a mean outside the normal model.

Inference should preserve pairing between arms, common market shocks across strategies, temporal dependence, and the actual selection rule. Uncertainty for a difference of dependent correlations must be estimated jointly. Estimating a persistence loading from near-unit correlations can be ill-conditioned; a rounded point solution is not evidence of precise identification. In the further special case $\kappa_O = \kappa_I$, the illustrative naive OOS dispersion is 0.434031, compared with 0.438 in sample. The observed OOS dispersion and its paired uncertainty would provide an additional model check, but no test result is claimed here.

The summary moments reported in this section are insufficient to perform those diagnostics or identify a live economic loss. No ranking of the deployment and same-tape probabilities of backtest overfitting follows from the correlation inequalities alone; those statistics require their own defined resampling experiment. These limitations do not negate the measured ranking comparisons, but restrict their attribution to persistent drift and their extrapolation to selection economics.

13.8 Reported planted-drift experiment

The reported intervention applies a causal momentum drift, of per-bar strength ε , to the price paths supplied to both arms, using the trailing return sign and resetting at the session open. We use ε , rather than μ , because the intervention is in per-bar price-drift units, whereas $\mu = \mathbb{E}[g]$ is a population-average score. Their relationship is not specified by the factor model. A predictable injected component also does not guarantee positive net performance after costs for a given strategy.

When a positive multiplier F_k is applied uniformly to the OHLC values of both versions of bar k , the aligned range comparison satisfies

$$H_k^{N'} - H_k^{E'} = F_k(H_k^N - H_k^E), \quad L_k^{E'} - L_k^{N'} = F_k(L_k^E - L_k^N),$$

so $\Delta_k' = F_k \Delta_k$. Relative contamination geometry is preserved within that bar, not the absolute dollar range difference. The transformation can nevertheless change strategy decisions, return volatilities, and score-factor relationships. The identity transformation at $\varepsilon = 0$ is a baseline check, not a proof that fidelity orderings are invariant to injection.

Reported outputs. Table 10 retains the three reported executable-OOS top-decile outputs from the 7,232-strategy, 60-name experiment over the 2014–2026 odd-lot period, net of the stated cost model. These are the experiment’s reported top-decile summary statistics; they are not equated to either an average of strategy Sharpes or a portfolio-return Sharpe without the corresponding aggregation definition. Consequently they are not used to calibrate Theorem 3 or to calculate a capital-scaled dollar loss.

Table 10: Reported executable-OOS top-decile statistics under the planted-drift intervention. The three settings show a higher clean-selected output, but none directly observes opposite signs between arms. Values are reproduced from the manuscript’s summary of `injection_sweep.json`; no additional run or aggregation convention is inferred.

Drift ε per bar	Clean-selected	Naive-selected	Difference
5×10^{-5}	−0.193	−0.215	0.022
1×10^{-4}	−0.067	−0.084	0.017
2×10^{-4}	0.287	0.260	0.027

Interpolation is not an observed deployment flip. Linear interpolation between the reported 10^{-4} and 2×10^{-4} settings gives descriptive zero crossings

$$\varepsilon_C^{\text{lin}} = 1.189266 \times 10^{-4}, \quad \varepsilon_N^{\text{lin}} = 1.244186 \times 10^{-4},$$

separated by 5.492051×10^{-6} . At their midpoint, 1.216726×10^{-4} , the interpolated outputs are approximately 0.009721 and -0.009446 . These numbers are calculated from the displayed rows, not additional experimental observations. Selected strategy sets can change discontinuously with the intervention, so their performance need not interpolate linearly. The reported sweep therefore suggests, but does not demonstrate, an opposite-sign interval.

Direct runs inside that interval, a documented aggregation rule, and paired uncertainty estimates are required before describing an observed portfolio deployment flip. Any extension of the sweep should distinguish exploratory runs from a prespecified confirmation. The three existing settings support only the descriptive result that the reported clean-selected statistic is higher at each tested strength. They do not independently validate the Gaussian magnitude or a dollar-cost estimate.

Remark 4 (Scope and dependence on the null). *The executable-gain lemma is conditional on the price-and-fill null. The fidelity formulas are finite-moment results with unrestricted mean, and their strict orderings require the stated covariance and noise restrictions. The distribution-free selection comparison requires conditional-mean sufficiency; its Gaussian counterpart adds a tail model to obtain a magnitude. Cleaning can raise or lower measured strategy performance, while a recurrent contamination component can help a strategy validate on a tape that does not represent its executable opportunity. The empirical ranking gaps are consistent with that mechanism, but the summary evidence does not uniquely distinguish it from transient score noise, correlated distortions, or heterogeneous denominators.*

14 Discussion

The result has a direct practical reading, and it is a cautionary one. A retail platform or a backtesting service that ships raw trade-only bars is shipping a backtest whose strategy selection is less faithful to live performance than it would be on executable bars: on the edge-free full universe, across twelve structural strategy families over the odd-lot era, in-sample selection on clean bars predicts the executable out-of-sample ranking with rank-IC 0.9987 against 0.9948 on naive bars, a gap of $+0.0039$ that is positive and significant under an effective- N cluster bootstrap and positive in every sub-period and de-duplication

construction tested, with odd lots supplying three-quarters of the underlying inflation. The naive backtest predicts its own contaminated out-of-sample better than the executable one, internally consistent and externally less valid. The dollar penalty of that infidelity cannot be realised on a naturally occurring edge-free universe, so we characterise it by proof: Theorem 3 gives the forfeited executable out-of-sample score in closed form, and Remark 3 evaluates it at the measured contamination parameters as an implied score-unit shortfall of about 0.015 at top-decile selection under a Gaussian benchmark. Converting that shortfall into a portfolio dollar loss requires executable portfolio return paths and a stated aggregation and annualisation convention, and is not identified from the shortfall alone; the planted-drift experiment of Section 13.8 is consistent with a penalty of this sign, finding the clean-selected book higher at every tested strength, without directly observing the opposite-sign interval or a dollar figure. The infidelity itself is removable with information the exchanges already publish. The fix is a data choice, not a modeling choice: build the marketable-reachable bar from prints executed at or inside the NBBO (the TRADE-NB construction), which keeps reachable inside-NBBO fills and drops only out-of-NBBO prints. Consuming the SIP official OHLC fields is only a partial fix: official OHLC drops odd lots and condition-coded prints but retains regular-way off-exchange “D” prints that printed outside the NBBO, last-sale eligible yet unreachable by a marketable order. The fully executable bar requires the NBBO anchoring, not the official OHLC fields. This orders three data tiers by which prints each retains: a light tier of raw bars, a pro tier of official-OHLC-style bars that drop odd lots and condition-coded prints, and a quant tier of full tick plus the NBBO-anchored executable bar. We measure the two endpoints, the light (naive) and quant (executable) arms; the pro tier’s intermediate standing is a consequence of the eligibility rules above, since official OHLC removes odd lots and suppressed conditions but still retains out-of-NBBO “D” prints, rather than a separately run selection arm. The honest-backtest literature has hardened the statistical side of backtest validity (deflation, PBO, purged cross-validation) while largely taking the input bars as given; this paper is the missing companion to Holden and Jacobsen on the data side, showing that the input bars are not given and that the difference between naive and executable bars is the difference between a faithful selection and a contaminated one. We are careful about what the full-universe result does and does not establish. It establishes that clean selection is more faithful than naive selection, in a universe where the right answer is to trade nothing. The full top-1000 universe contains no naturally occurring profitable strategy, so the positive claim that clean bars let a profitable strategy survive while naive bars degrade it is characterised mathematically, by Theorem 3 and Corollary 2, and probed by the planted-drift experiment of Section 13.8, which finds the clean-selected book higher at every tested strength short of a directly observed flip. The mechanism builds on the spread-direction effect Corwin and Schultz identified, but the practitioner-facing harm is sharper than uniform optimism: the bar range is sign-definite (Proposition 1), yet the backtest distortion it produces is two-sided and strategy-specific, inflating most strategies in-sample through phantom breakouts and deflating a minority through spurious stop-outs, so it scrambles the in-sample ranking rather than shifting it, and a scrambled ranking is what fails to select live.

15 Conclusion and Limitations

In one sentence: on the full US equity top-1000 universe over the odd-lot era, where the universe is edge-free by construction and nothing is profitable out-of-sample, selecting strategies across twelve structural families on executable, NBBO-anchored bars predicts the executable out-of-sample ranking with rank-IC 0.9987, better than selecting on naive trade-only bars (0.9948), a fidelity gap of +0.0039 that is positive and significant under an effective- N cluster bootstrap (+0.020 on the uncorrelated subset), positive in every sub-period year and de-duplication construction, while the identical universe on CME futures, which carry no phantom prints, shows a gap of zero, because the apparent in-sample edge in the naive equity bars is manufactured by untradeable prints. The claim is one-sided and deliberately a null on profit: every result we report on the natural tape is one that an edge-free universe can carry, and every claim that would require a genuine edge to be meaningful is proved (Theorem 3, Corollary 2) and probed by a planted-drift experiment (Section 13.8), never asserted from a naturally occurring profit. Cleaning removes a false in-sample edge and restores selection fidelity; it does not manufacture an edge where none exists, and it cannot, since removing prints removes apparent performance rather than adding real performance. That the universe is edge-free is what makes the claim literally true here: there is no real edge for cleaning to find, and naive selection is still measurably less faithful.

The dominant remaining gap is coverage, and we name it first because it is the one threat that bounds the whole paper.

(i) Coverage is the full addressable universe; the residual is the genuine-edge case.

The addressable universe is the full US equity tape (top-1000 names by dollar volume, 2007 to 2026) plus matched CME futures, and this paper runs on it: 998 names over 1.46 billion bars, with the strategy experiment on the odd-lot era 2014–2026. Coverage is therefore not the residual gap. The residual is that the naturally occurring universe is edge-free, so the one result no measurement on that tape can deliver is the positive commercial claim, that clean bars let a genuinely profitable strategy survive while naive bars degrade it, because a naturally occurring edge-free universe contains no profitable strategy to survive. We carry that result and the dollar penalty as proved consequences (Theorem 3, Corollary 2, Remark 3) and, for the genuine-edge case, probe them with the planted-drift experiment of Section 13.8, which finds the clean-selected book higher at every tested strength short of a directly observed flip; we flag which is which wherever they appear.

(ii) The universe is edge-free, so the headline is a fidelity result, not profit.

No strategy is profitable out-of-sample (best clean-out-of-sample Sharpe -0.13 across the corpus and -0.30 on the uncorrelated subset, deflated Sharpe zero for both arms). The realized top-decile selection-value gap is zero on this edge-free universe, both arms selecting the same least-bad strategies, so here the penalty is carried by Theorem 3 as a score-unit shortfall; the planted-drift experiment of Section 13.8 is consistent with it but does not identify a portfolio dollar loss.

(iii) Residual spread and optimistic exits. The executable bar is still wider than the true round-trippable price by the spread component (Corwin and Schultz; EDGE), which no condition-code filter removes; the headline compares naive against executable selection,

both of which retain that residual. The cost model is also optimistic on exits: a stop-loss fills at the stop price charged only the modeled half-spread, with no gap or impact slippage, which flatters the fast-exit breakout families most. Because the naive and clean arms trade at parity ($1.006\times$ the fills), the residual applies near-identically to both and enters the gap only at second order. A cost-multiplier stress confirms this: scaling the per-fill friction by $2\times$, $3\times$, and $5\times$ leaves the fidelity gap positive and significant, near-invariant at roughly $+0.003$ (Section 10). What the model does not resolve is within-bar queue position and market impact beyond the half-spread, so the exit cost of a fast-turnover strategy is understated in absolute terms; this affects both arms alike and is a bound on the realized-dollar figures, not on the fidelity gap.

(iv) Inference base. The unit of analysis is the strategy (a cross-symbol portfolio over 998 names), and the 7,232 strategies are highly correlated, so an i.i.d. bootstrap over them overstates precision. We price significance instead at the effective number of independent bets: each strategy is assigned to its most-correlated de-duplicated representative, forming ≈ 20 correlation blocks, and the block bootstrap gives the gap $[0.0025, 0.0093]$, $p < 0.001$ (Table 5); the i.i.d.-over-strategies diagnostic is far tighter and is reported only to show the overstatement. The gap is positive in every one of the thirteen sub-period years and every de-duplication threshold. The rank-IC level is partly mechanical (cost and turnover persist across windows, so relative badness ranks stably even with no edge), so the decisive quantity is the naive-versus-clean gap, not the absolute IC.

(v) Strategy-family scope, and no attributed signal mechanism. The corpus spans twelve structural families (Section 5.4). We do not attribute the contamination to any single signal mechanism: the per-family gaps are heterogeneous with overlapping intervals, and the family-clustered gap ($+0.020$, 95% interval $[0, 0.062]$, $p = 0.12$) is consistent with the headline but does not isolate a family. The headline gap does not depend on the partition. The families tested already span trend, breakout, momentum, oscillator, and mean-reverting shapes.

(vi) The controls bound and sharpen the headline; they do not replace it. The source decomposition localizing three-quarters of the inflation to odd lots (Section 7), the two-arm CME futures negative control (Section 11, gap -0.00011 against the equity $+0.0039$), and the random-print-drop placebo (Section 12) together establish that the effect is specific to removing untradeable equity prints and is not a pipeline artifact or a “fewer prints” effect. The tick-resolved intrabar measurement (Section 9) shows the intrabar channel is small and near pessimistic. The point-in-time universe, including distressed delistings by construction, strengthens rather than weakens the gap (Section 10).

(vii) The executable bar is not the official OHLC fields. Official OHLC applies the SIP’s suppression logic (it drops odd lots and condition-coded prints), but it retains regular-way off-exchange “D” prints that printed outside the NBBO, last-sale eligible yet unreachable by a marketable order, so consuming official OHLC is only a partial clean. The executable bar anchors to the NBBO (the TRADE-NB construction, trades at or within the NBBO), keeping reachable inside-NBBO prints and excluding only the out-of-NBBO ones. The contribution is measuring what the naive bar costs a backtest and specifying the marketable-reachable bar that fixes it, which is a distinct object from official OHLC rather

than a subset of it: it keeps reachable inside-NBBO prints that official OHLC drops and excludes out-of-NBBO prints that official OHLC retains.

The seven non-negotiable gates, where a reader can check them. No look-ahead: every filter uses only the print-time condition code (Section 5.1), the family port is verified causal (the higher-timeframe trend on a series truncated at t is bit-identical to the full series at the last common index, the session gate is a pure clock, fills are at $t+1$, Section 5.4), and the walk-forward procedure itself is pollute-and-verify tested for leakage (Section 5.5). No costless backtest: a modeled time-of-day half-spread plus per-fill commission is charged round-trip and stress-tested at $2\times$, $3\times$, and $5\times$ (Section 10). Real data: the Algoseek consolidated tape and CME futures tape, with synthetic appearing only in the placebo, the no-leak pollution test, and the labelled known-answer edge injection of Section 13.8. A null per claim: the dose-response, the random-drop placebo, the two-arm futures control, and the executable-bar structural null, each run through the identical pipeline (Sections 8, 11, 12). Mandatory walk-forward: rolling IS of 126 days and OOS of 21 days, not a single split (Section 5.4). Survivorship owned: the universe is point-in-time and includes delisted names by construction, so failed and distressed tapes are already in the corpus on the dates they traded, which makes the headline conservative rather than optimistic (Section 10). Headline scoped: every claim is bounded to the tested universe, period, and family set, with the genuine-edge case foregrounded above.

A Condition-Code to OHLC-Eligibility Mapping

This appendix expands the material subset of Table 1 into the per-field eligibility (last sale, high/low, and volume) with the suppression-wins rule applied field by field; conditions not listed inherit the regular-way base case (eligible for last, high, low, and volume). Source: CTS Pillar Input/Output Specification, the UTP Sale Condition Matrix, the FINRA Trade Report Modifiers chart, and the Algoseek Trade-Only guide.

B Reproducibility and Data Availability

Every headline number in this paper rests on Algoseek US equity tapes that are licensed and non-redistributable: they are obtained under provider terms and read from a local data directory with credentials taken from environment variables, and they cannot be regenerated from the public repository alone. The repository therefore ships the engine and its correctness guarantees (the bar-construction filter, the walk-forward harness, the no-look-ahead pollution test, the cost model) exercised on bundled fixtures, plus the committed per-run result JSONs from which the figures render directly. Determinism is by fixed seed (12345) with a 5,000-replication bootstrap. The primary endpoint, the rank-IC, is the Spearman rank correlation (SciPy `spearmanr`) between the per-strategy in-sample and executable-out-of-sample Sharpe, computed once over the cross-symbol strategy corpus (all 7,232 strategies for the point estimate), not per-symbol then averaged; the deduplication threshold τ is a separate quantity, the maximum-absolute-Pearson-correlation cutoff on clean-out-of-sample daily returns used in the greedy de-correlation step, and is neither Kendall's tau nor a rank correlation. The universe is the full top-1000 by dollar volume (998 names; prevalence and decomposition over 2007–2026, 1,464,795,700 RTH bars; the strategy experiment over the odd-lot era 2014-01-02 to 2026-07-24, 3,158 trading

Table 11: Per-field condition-code to OHLC eligibility. “Last”, “High/Low”, and “Volume” are the consolidated-tape update eligibilities; a single trade can be eligible for one field and not another (for example, prior-reference and out-of-sequence prints update the high and low but not the last sale). Conditions not listed inherit the regular-way base case. Source: CTS Pillar and UTP specifications; FINRA Trade Report Modifiers; Algoseek Trade-Only guide.

Condition	Flag	Last	High/Low	Volume	Note
Regular-way trade	(none)	Yes	Yes	Yes	The eligible base case.
Intermarket sweep (ISO)	tIntermarketSweep	Yes	Yes	Yes	Eligible despite multi-price execution.
Opening/closing auction trade (“O”/“6”)	tOfficialOpen / tOfficialClose	Yes	Yes	Yes	Auction execution.
Official open/close message (“Q”/“M”)	(indicative)	n/a	n/a	n/a	Indicative price message, not a trade; excluded by the bar builder.
Odd lot (“I”)	tOddLot	No	No	Yes	Volume-eligible only since 2013-12-09.
Average price (“W”)	tAveragePrice	No	No	Yes	Computed average, not a market print.
Prior reference (“P”)	tPriorReferencePrice	No	Yes	Yes	Updates high and low, not last.
Derivatively / benchmark	tDerivativelyPriced	No	Yes	Yes	Updates high and low, not last.
Out of sequence (“Z”)	tOutOfSequence	No	Yes	Yes	Updates high and low, not last.
Form-T / ext. hours (“T”)	tFormT	No	No	Yes	Pre/post-market; excluded from regular-session OHLC.
Cash / next-day	tCash / tNextDay	No	No	Yes	Non-regular-way settlement.
Off-exchange “D”, regular	code “D”, no condition	Yes	Yes	Yes	Last-sale eligible; the executable bar keeps only the at-or-inside-NBBO subset.
Off-exchange “D”, flagged	“D” with .P/W/late/QCT/Form-T	per flag	per flag	Yes	Eligibility follows the attached condition.

days); the framework was frozen on an eight-name stratified development panel (NVDA, AAPL, TSLA, SPY, QQQ, SQQQ, BAC, F; 2021–2023) before the full universe was analysed. Every headline number is reproducible from a committed result JSON, and the figures render from the committed figure set. Engineering follows the project rules: zero-disk S3 streaming, processes not threads, Numba or Rust hot loops, parquet-native output, capped disk-heavy writers, heartbeat-monitored long runs, and pinned BLAS threads.

C Results Inventory

A complete inventory of every reported number with its producing script and result JSON accompanies the repository, so each tagged quantity in the body maps to one artifact.

D Figure Provenance

For each figure, the result JSON it renders from and the script that produced it accompany the repository.

E Pre-Registration Hash

The analysis plan was fixed in code before the full-universe compute. The block naming the rank-IC gap as the primary endpoint, together with the de-duplication threshold, the ranking key, the bootstrap replication count and the seed, is byte-identical in the research repository across 2026-06-26 and 2026-06-30, and the full-universe run started on 2026-07-23 and completed on 2026-07-26. That plan is released with the replication code as `preregistration/analysis-plan-locked.md` and its SHA-256 is recorded beside it in `preregistration/LOCK_HASHES.md`, so a third party can recompute the digest against the released file. It covers the experiment design of Section 5.4, the metric definitions of Section 5.5, and the stratified-panel definition where the tick budget binds. The released file is a transcription of the frozen plan rather than a document that itself existed on the freeze date; the ordering evidence is the repository history summarised above. Five analyses were added after the freeze and are labelled in the plan as confirmatory rather than pre-registered. The largest is the effective- N block bootstrap, which replaced the symbol-clustered bootstrap of the locked plan because the unit of analysis is a cross-symbol portfolio, leaving the symbol cluster undefined; it is the more conservative choice and the locked alternative is reported as the anticonservative diagnostic in Section 8. One planned analysis, PBO via CSCV, is not reported: the de-duplicated corpus is too small for the combinatorially symmetric split to run.

References

- [1] Ait-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of microstructure noise. *Review of Financial Studies* 18(2), 351–416.
- [2] Ardia, D., Guidotti, E., and Kroencke, T. A. (2024). Efficient estimation of bid-ask spreads from open, high, low, and close prices. *Journal of Financial Economics* 161, 103916.
- [3] Bailey, D. H., Borwein, J. M., Lopez de Prado, M., and Zhu, Q. J. (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.
- [4] Bailey, D. H., Borwein, J. M., Lopez de Prado, M., and Zhu, Q. J. (2017). The probability of backtest overfitting. *Journal of Computational Finance* 20(4), 39–69.
- [5] Bailey, D. H., and Lopez 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.
- [6] Bandi, F. M., and Russell, J. R. (2008). Microstructure noise, realized variance, and optimal sampling. *Review of Economic Studies* 75(2), 339–369.

- [7] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* 76(6), 1481–1536.
- [8] Bartlett, R. P., McCrary, J., and O’Hara, M. (2023). The market inside the market: Odd-lot quotes. *Review of Financial Studies* 38(3), 661–701.
- [9] Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics* 24(2), 265–272.
- [10] Concretum Group (2024). How to detect and fix problems in intraday market data. `concretum_tests` Python library and accompanying note. <https://concretumgroup.com/coding/>.
- [11] Corwin, S. A., and Schultz, P. (2012). A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance* 67(2), 719–760.
- [12] Glosten, L. R., and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14(1), 71–100.
- [13] Hansen, P. R., and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics* 24(2), 127–161.
- [14] Holden, C. W., and Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance* 69(4), 1747–1785.
- [15] Lee, C. M. C., and Ready, M. J. (1991). Inferring trade direction from intraday data. *Journal of Finance* 46(2), 733–746.
- [16] Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal* 58(4), 36–52.
- [17] Low, R. K. Y., Maier-Paape, S., and Platen, A. (2016). Correctness of backtest engines. Working paper, arXiv:1509.08248.
- [18] Maier-Paape, S., and Platen, A. (2014). Backtest of trading systems on candle charts. Working paper, arXiv:1412.5558; *IFTA Journal* 2016, 10–17.
- [19] Mesfin, M. (2026). Structural limits of OHLCV-based intraday signals in MNQ futures: A systematic falsification study. Working paper, arXiv:2605.04004.
- [20] O’Hara, M., Yao, C., and Ye, M. (2014). What’s not there: Odd lots and market data. *Journal of Finance* 69(5), 2199–2236.
- [21] Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* 39(4), 1127–1139.
- [22] Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology* 15(1), 72–101.

- [23] Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)* 23(1), 223–229.
- [24] Zhang, L., Mykland, P. A., and Ait-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100(472), 1394–1411.
- [25] US Securities and Exchange Commission. Regulation NMS, Release 34-51808 (2005); Market Data Infrastructure Rule, Release 34-90610 (2020); Reg NMS tick-size, access-fee, and round-lot amendments, Release 34-101070 (2024); odd-lot tape-reporting order, Release 34-70794 (2013).