Prevalence Determines Precision: Silent Contamination in Detector-Defined Datasets

Jia Huang[1], Yankai Wan[2],Yangjun Ou[3]

[1]Guanghua School of Management, Peking University
[2]College of Artificial Intelligence, Jilin University
[3]School of Mathematical Sciences, Peking University

Abstract

A large share of machine learning datasets are not observed but *constructed*: a detector, heuristic, or model is run over a pool of candidates, and whatever it accepts becomes the dataset. Weak and distant supervision, pseudo-labelling, event extraction, and most anomaly-detection benchmarks all have this form. The precision of the resulting dataset is not a property of the detector. It is governed by the prevalence π of true positives in the pool the detector is deployed on, through elementary Bayes. This is textbook, and it is nonetheless almost never measured end-to-end, for a structural reason: from inside a detector-defined dataset the false positives are unidentifiable, so the quantity that matters cannot be estimated by the people who need it. We report a case in which it *can* be measured. For the same instrument and period we hold both a detector-defined event dataset and an independent official index that reveals, for every detected item, whether it is real. One detector, three candidate pools, three datasets: phantom rates of 81.7%, 9.0%, and 0.0%. A practitioner transferring precision from the two high-π pools to the low-π pool predicts 0.955 against a measured 0.183, an error of $+422\%$; the Bayes expression predicts within 3.3% across all three pools. Beyond confirming the mechanism we report three findings that we believe are new. First, the detected dataset's response curve is an *exact* convex combination of a true-event and a phantom component (identity residual 1.1×10^{-16}), with phantoms outnumbering true events 473 to 308 — contamination is not noise added to signal but a second signal with its own shape, inherited from the detector's acceptance rule. Second, the *direction* of contamination is a property of the estimator, not of the data: on identical windows, one statistic shows contamination diluting the effect and another shows it inflating the effect, because the second statistic's denominator is itself contaminated. Third, a normalisation in common use turns the estimator into a mean of ratios whose expectation does not exist; on the same 335 events it returns 0.40 where the well-defined estimator returns 0.10.

1 Introduction

Consider how a dataset gets built when nobody is willing to label 10^6 items by hand. A rule is written — a regular expression, a knowledge-base join, a threshold on a model's confidence, a statistical scan — and applied to a pool of candidates. What the rule accepts becomes the dataset. Downstream, the accepted items are treated as observations. This is the shape of distant supervision (Mintz et al., 2009), of weak supervision (Ratner et al., 2017), of pseudo-labelling, and of the great majority of event-detection and anomaly benchmarks.

The precision of such a dataset obeys

$$\text{precision} \;=\; \frac{\pi\,\mathrm{TPR}}{\pi\,\mathrm{TPR} + (1-\pi)\,\mathrm{FPR}}, \tag{1}$$

where π is the fraction of true positives in the pool the detector was run on. Precision is therefore a property of the *deployment pool*. What transfers across pools is the pair $(\mathrm{TPR}, \mathrm{FPR})$; precision does not. A detector validated where $\pi \approx 1$ — and at $\pi \approx 1$ even a badly broken detector looks flawless — and then redeployed where $\pi \ll 1$ imports false positives at a rate its validation never measured.

None of this is new. Bar-Hillel (1980) named the reasoning error; Hand (2009) and Saito & Rehmsmeier (2015) state the consequence for classifier evaluation directly. What is missing is measurement. The obstacle is structural

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rather than sociological: inside a detector-defined dataset the false positives are exactly the items one cannot identify. If they could be identified, the detector would not have accepted them. Practitioners therefore validate on whatever labelled subset they have — which is, almost by construction, a high-π subset — and deploy elsewhere.

This paper. We report a setting where the phantoms *are* identifiable. For E-mini S&P 500 futures over 2010–2026 we hold two independent event indices over the same price series: a detector-defined set produced by a magnitude-threshold scan, and an official index (CPI, non-farm payrolls, and FOMC statement releases) published by the institutions that generate the events. Every detected item can be classified. Moreover the detector was run on three structurally different candidate pools — the Federal Reserve's own meeting calendar, the set of first Fridays, and business days 6–16 of each month — giving three deployments of one detector at $\pi \in \{1.00,\ 0.82,\ 0.10\}$ within a single dataset.

We use this to do three things a synthetic study cannot. We measure (1) against reality rather than assuming it. We decompose a real contaminated estimate into its true and phantom components exactly, and characterise what the phantom component looks like. And we show that the contamination's effect on a downstream scientific conclusion is not even sign-stable across reasonable choices of estimator.

Contributions.

1. **An end-to-end measurement of prevalence-driven contamination** with recoverable ground truth: one detector, three pools, phantom rates 81.7% / 9.0% / 0.0%; naive cross-pool precision transfer errs by $+422\%$ while (1) holds to 3.3% (§5.1–§5.2).
2. **An exact mixture decomposition** showing that a contaminated estimate is a convex combination of a true-event and a phantom estimate, verified to machine precision on real data, with the phantom component reproducible from the detector's acceptance rule applied to pure noise (§5.3).
3. **Estimator-dependence of the contamination direction.** On identical windows, contamination dilutes under a denominator-free statistic and inflates under a terminally normalised one. We trace this to contamination of the estimator's own scale (§5.4).
4. **A non-integrable estimator in common use.** Per-item normalisation by a signed reference return yields a mean of ratios with no finite expectation. Empirically it returns 0.40 where the equal-weight estimator returns 0.10 on the same events (§5.5).
5. **A seven-rule protocol** (§7) and a six-accident appendix (App. A) documenting the mechanism operating on this paper's own pipeline.

2 Setup

Observed and detected indices. Let p_t be a series and let T^{obs} be an *observed* index published by an external authority. Let A be a *detector*: any deterministic rule mapping a candidate window to accept/reject. The *candidate pool* C is the set of windows at which A is evaluated. The *detected index* is $T^{\mathrm{det}} = A(C)$. Users of T^{det} typically see only $|T^{\mathrm{det}}|$.

Five quantities. We argue five must be reported alongside any detector-defined dataset:

$$\pi = \frac{|T^{\mathrm{obs}} \cap C|}{|C|} \qquad \text{prevalence in the deployment pool}$$

$$\mathrm{TPR} = \frac{|T^{\mathrm{obs}} \cap T^{\mathrm{det}}|}{|T^{\mathrm{obs}} \cap C|}, \qquad \mathrm{FPR} = \frac{|T^{\mathrm{det}} \setminus T^{\mathrm{obs}}|}{|C \setminus T^{\mathrm{obs}}|} \qquad \text{detector operating point}$$

$$c = \frac{n_P}{n_E + n_P} \qquad \text{phantom mass fraction}$$

$$\rho = w_P / w_E \qquad \text{phantom-to-true intensity ratio}$$

where $n_E = |T^{\mathrm{obs}} \cap T^{\mathrm{det}}|$, $n_P = |T^{\mathrm{det}} \setminus T^{\mathrm{obs}}|$, and w_E, w_P are the mean magnitudes of the two subsets under whatever scale the downstream estimator uses. Precision equals $1 - c$, and by Bayes it equals (1).

The last two are the ones normally omitted even by careful authors. c says how much of the dataset is fabricated; ρ says how loud the fabricated part is relative to the real part. §5.3 shows they jointly determine the damage, and that c alone does not.

Response curves. For an index T with anchor sign $s_i = \operatorname{sign} r_i(\tau_{\mathrm{ref}})$, define

$$S(\tau) = \frac{1}{n}\sum_i s_i\, r_i(\tau), \qquad D_{\mathrm{ref}}(\tau) = \frac{S(\tau)}{\frac{1}{n}\sum_i |r_i(\tau_{\mathrm{ref}})|}, \qquad \widetilde{D}(\tau) = \frac{1}{n}\sum_i \frac{r_i(\tau)}{r_i(\tau_{\mathrm{ref}})}, \tag{2}$$

where $r_i(\tau)$ is the log return from t_i to $t_i + \tau$. S is denominator-free and reported in basis points. D normalises by a pooled scale. $\widetilde{D}$ normalises per item; it looks like a natural way to make heterogeneous events comparable, and §5.5 shows it is not an estimator at all. We use $\tau_{\mathrm{ref}} = 1800\,\mathrm{s}$ throughout the main text and report the legacy pipeline's $\tau_{\mathrm{ref}} = 7200\,\mathrm{s}$ convention in Appendix D.

3 Theory

3.1 Precision does not transfer; $(\mathrm{TPR}, \mathrm{FPR})$ does

Calibrating a detector means choosing a threshold, which fixes an operating point $(\mathrm{TPR}, \mathrm{FPR})$. Those are conditional on the signal and noise distributions, not on π. Precision is then determined by (1) in any pool where π is known. The failure mode is not subtle once stated, but note what makes it invisible in practice: at $\pi \approx 1$, precision is near 1 for *any* FPR, so a validation at high π carries essentially no information about FPR — the very quantity that governs behaviour at low π. High-prevalence validation is not weak evidence about low-prevalence deployment; it is close to no evidence.

3.2 The mixture identity

Let E, P partition a detected index into true events and phantoms, with intensities w_E, w_P under the estimator's own scale. Because the anchor sign is defined identically on both subsets, for every τ

$$D^{\mathrm{det}}(\tau) \;=\; \frac{(1-c)\,w_E\,D_E(\tau) \;+\; c\,w_P\,D_P(\tau)}{(1-c)\,w_E + c\,w_P} \;=\; \frac{(1-c)\,D_E(\tau) + c\,\rho\,D_P(\tau)}{(1-c) + c\,\rho}. \tag{3}$$

This is an identity, not a model: it holds for any detector, any τ, any data. Its content is diagnostic. A dataset owner who knows (c, ρ, D_E, D_P) knows exactly what their published curve measures. A user holding only D^{det} cannot invert it, which is precisely the difficulty: contamination is not visible from inside the contaminated dataset.

Two consequences. First, contamination is *not* additive noise. D_P has a shape, and when the detector selects on the same quantity that defines the phenomenon, that shape resembles the phenomenon. Second, the weights involve ρ, so a dataset with high precision but loud phantoms can be more damaged than one with low precision and quiet phantoms.

3.3 The phantom null is a selection-conditioned null

Phantoms are, by construction, windows accepted by the rule. Under a pure-noise null, conditioning on a large move at the acceptance horizon induces structure in any statistic anchored near that horizon — not because the noise contains an event, but because the conditioning event is itself part of the statistic. The correct null for D_P is therefore not "flat" and not the unconditional random walk, but the *detector's transfer function*: noise passed through the same acceptance rule and aggregated by the same estimator. We construct this null in §5.3.

3.4 A normalisation with no finite expectation

Proposition 1. *Let r_{ref} have a density f that is continuous and strictly positive in a neighbourhood of 0. Then $\mathbb{E}\big[|r(\tau)/r_{\mathrm{ref}}|\big] = \infty$, and the estimator $\widetilde{D}(\tau) = \frac{1}{n}\sum_i r_i(\tau)/r_i(\tau_{\mathrm{ref}})$ of (2) has no finite expectation.*

Proof sketch. $\mathbb{E}|1/r_{\mathrm{ref}}| = \int f(x)/|x|\,dx$ diverges logarithmically at the origin whenever $f(0) > 0$. Conditional on $r(\tau)$ having non-degenerate conditional mean near $r_{\mathrm{ref}} = 0$, the same divergence carries to the ratio. □

The sample mean of $\widetilde{D}$ is finite for any finite n, so the pathology is silent: one obtains a number, it looks stable across reruns on the same data, and it does not converge to anything as n grows. It is dominated by the items with the smallest $|r_{\mathrm{ref}}|$ — which, in an event dataset, are the items where nothing happened. §5.5 measures the size of the resulting artefact.

4 A dataset with recoverable phantoms

Price series. E-mini S&P 500 futures (front-month continuous, `ES.c.0`), best-bid-offer at 1-second resolution from a CME-derived vendor feed, in windows of $t_0 \pm 2\,\mathrm{h}$. Total spend $2.04.

Official index. 555 announcement windows over 2010–2026: 212 CPI and 203 non-farm payroll releases (release dates from the Federal Reserve Bank of St. Louis release calendar; release times from the statutory 08:30 ET convention), and 140 FOMC statements (dates and per-statement release times verified against the Federal Reserve's own press-release pages; 14:15 ET before March 2013, 14:00 ET after, with seven unscheduled statements verified individually).

Detected index. A magnitude-threshold scan of the same price series: a window is accepted when the 2-minute move exceeds ten times a per-second standard deviation estimated on the pre-window. Full specification in Appendix E. The scan was run separately on three candidate pools, which is what makes this dataset useful:

Pool	Definition	$\lvert C\rvert$
FOMC	the Federal Reserve's published meeting calendar	140
NFP	the first Friday of each month	200
CPI	business days 6–16 of each month	1569

Screens. Applied identically to every label set before any comparison: a pre-window futures-roll screen (any 1-second log return above 20 bp strictly *before* t_0; post-t_0 jumps are deliberately excluded so the screen cannot delete announcements — see App. A, Accident 4) and a degraded-quote screen. 415 official windows survive for the terminal-anchor analysis and 417 for the 30-minute-anchor analysis, all from ∼2013 onward; the vendor's 1-second book history does not extend reliably before then.

Why this is a good testbed and where it is limited. The two indices are generated by causally independent processes: one by statistical agencies and a central bank, the other by a threshold on prices. Neither was constructed with the other in view. That independence is what makes phantom identification credible. The limitations are equally clear and we state them here rather than only in §8: one instrument, one detector family, one class of official index.

5 Results

5.1 One detector, three pools, three datasets

Table 1 is the paper's central measurement. Three observations.

The spread is the finding. One detector, one price series, one implementation, three pools — and three qualitatively different datasets. On the Fed's calendar the detector is perfect. On business days 6–16, four out of every five accepted items never happened.

Aggregation conceals it. The pooled phantom rate is 59%, which reads as a mediocre-but-usable detector. Per-pool, one of the three datasets is majority-synthetic. A practitioner who reports a single precision figure for a detector deployed across heterogeneous pools has reported a number that describes none of the deployments.

Table 1: The same detector deployed on three candidate pools. "Phantom rate" is n_P/n_{legacy}, i.e. one minus the precision of the detector *in that pool*. Counts are raw detections before quality screens; the screened counts used in §5.3 are $n_E = 308$, $n_P = 473$.

Pool	$\lvert C\rvert$	π	n_{official}	n_{legacy}	n_E	n_P	phantom rate
FOMC	140	1.000	140	126	126	0	**0.0%**
NFP	200	0.820	203	122	111	11	**9.0%**
CPI	1569	0.105	212	596	109	487	**81.7%**
All	1909	—	555	844	346	498	59.0%

Phantoms dominate in count. On the screened sample, $n_P = 473$ against $n_E = 308$: $c = 0.606$. A majority of the published event list consists of events that did not occur.

5.2 Cross-pool transfer: naive $+422\%$, Bayes -3.3%

The three pools give a natural calibration/deployment split. Calibrate on the two high-prevalence pools (FOMC, $\pi = 1.00$; NFP, $\pi = 0.82$), deploy on the low-prevalence one (CPI, $\pi = 0.105$).

A practitioner who treats precision as a detector property transfers the mean of the observed precisions, $\frac{1}{2}(1.000 + 0.910) = 0.955$. The measured CPI precision is 0.183: an error of $+\mathbf{422}\%$. The Bayes expression (1), using each pool's own π, that pool's TPR, and the detector's analytic $\text{FPR} = 0.361$ (derived in App. A, Accident 1, and confirmed by Monte Carlo to within 2%), predicts 0.177 against 0.183 measured: an error of $-\mathbf{3.3}\%$. The corresponding errors are -1.6% for NFP and 0% for FOMC. Figure 1 shows all three.

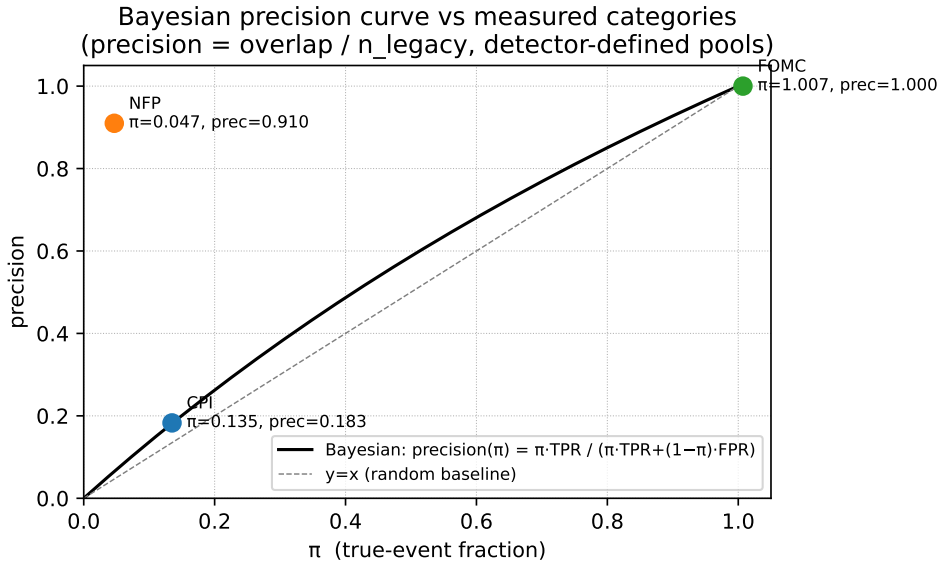

Figure 1: Precision as a function of pool prevalence. Solid and dashed black: (1) at the two measured TPR values, with $\text{FPR} = 0.361$ fixed. Dotted grey: the naive belief that precision is a property of the detector. Three measured deployments of one detector fall on the curve. The red marker is the naive transfer from the two high-π pools to the low-π pool; the dashed arrow is the resulting error.

We stress what is and is not being claimed. (1) is elementary and we claim no novelty for it. The claim is that its consequences are large, are realised in a real dataset built by competent people, and are not detectable from inside that dataset. The naive transfer is not a straw man: it is what "we validated the detector and got 95% precision" means when the validation pool and the deployment pool differ in prevalence by an order of magnitude.

5.3 Mixture decomposition: the contaminated curve, exactly

Applying (3) to the screened sample:

$$n_E = 308, \quad n_P = 473, \quad c = 0.606, \quad w_E = 6.38\,\text{bp}, \quad w_P = 3.76\,\text{bp}, \quad \rho = 0.590,$$

and the identity check

$$\max_{\tau} \left| D^{\text{det}}_{\text{pred}}(\tau) - D^{\text{det}}_{\text{obs}}(\tau) \right| \;=\; 1.1 \times 10^{-16},$$

i.e. exact to machine precision (Figure 2, right). This is first a verification that the bookkeeping is sound, and second a statement with content: the published curve of the detected dataset contains exactly the information in (3) and nothing else. There is no residual to interpret.

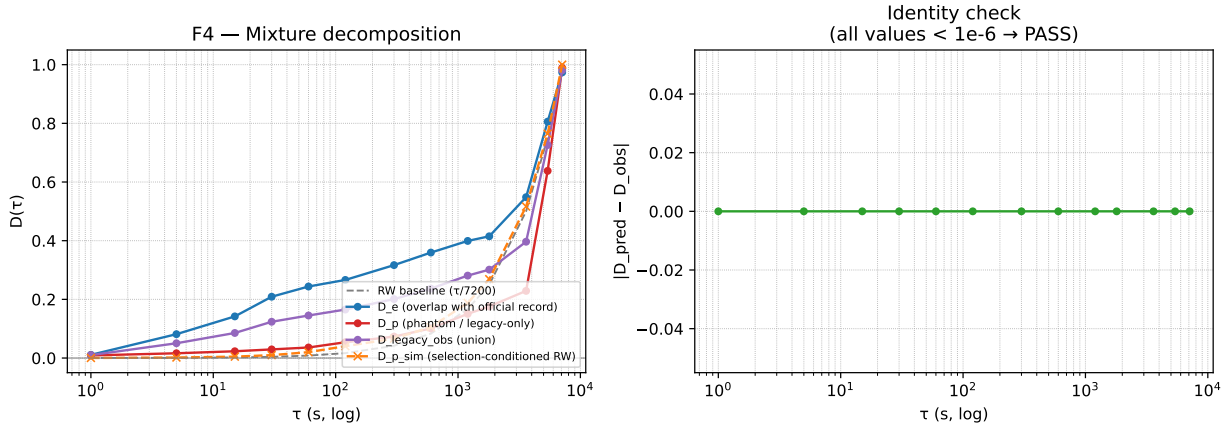

Figure 2: **Left:** the detected dataset's curve decomposed into its true-event component D_E, its phantom component D_P, and the exact mixture, with the unconditional random-walk baseline and the selection-conditioned null for D_P. **Right:** identity (3) verified to machine precision.

What the phantom component is. D_P is not flat and not the unconditional random walk. It rises with τ, qualitatively like D_E. Compared against the selection-conditioned null of §3 — pure noise passed through the detector's exact acceptance rule and the same estimator — it is reproduced in shape but not in level (RMSE 0.092 over $\tau \leq 600$ s, against D_P values in the range 0.02–0.07; the null sits systematically low). We therefore state the weaker claim that the data support: the phantom component's *shape* is generated by the acceptance rule rather than by event content, while its level is not fully accounted for by our null. We regard the residual as an open question rather than evidence of event content in the phantoms.

Why ρ matters. Here $\rho = 0.59$: phantoms are quieter than true events, so their mixture weight is below their count share. A dataset with the same c but $\rho > 1$ would be substantially more damaged. Reporting precision alone ($= 1 - c$) does not distinguish these cases.

5.4 The direction of contamination is a property of the estimator

This is the finding we expect to generalise furthest beyond the domain.

Under the terminally normalised statistic used by the original pipeline, contamination *inflates*: at $\tau = 5$ s, legacy 0.082 versus official 0.068; at 300 s, 0.317 versus 0.271. The contaminated dataset looks like it has *more* structure than the truth. Under the denominator-free statistic $S(\tau)$ on the same windows, the ordering reverses: official 0.79 bp $>$ legacy $0.48 >$ phantom-only 0.18 at $\tau = 5$ s, and $2.53 > 1.65 > 0.83$ at 300 s. Contamination *dilutes*. Figure 3 shows both.

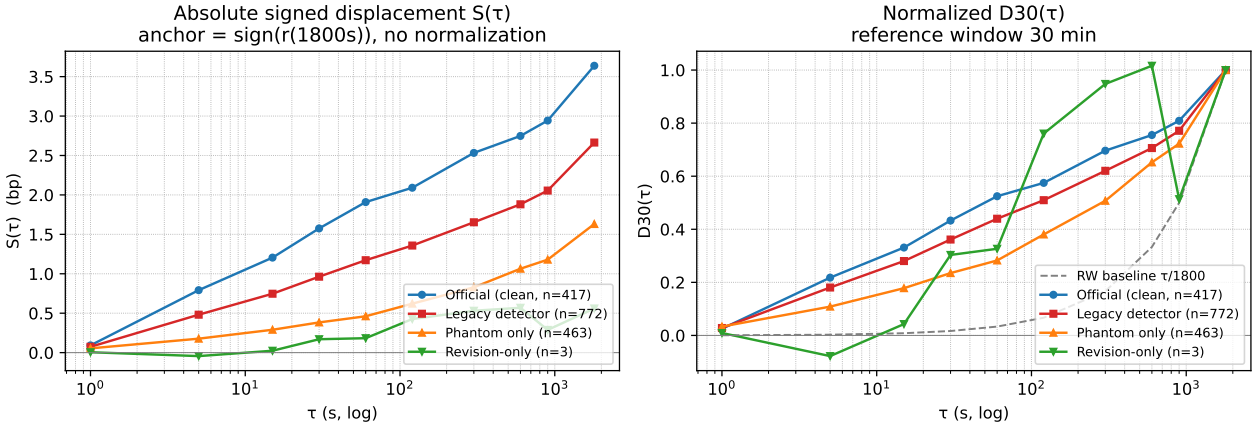

Figure 3: Same price windows, two label sets, two estimators. **Left:** denominator-free signed displacement $S(\tau)$ in bp. **Right:** normalised $D_{30}(\tau)$ with its random-walk baseline $\tau/1800$. Official n=417, detected n=772, phantom-only n=463. *The revision-only control is plotted at* n=3 *and is not interpretable*; only 3 of the 17 calendar-identified control windows have price coverage (App. C). Under terminal normalisation (App. D) the official/detected ordering is reversed relative to the left panel.

The mechanism is (3) together with the denominator. S has no denominator, so the mixture simply pulls the true curve toward the phantom curve: dilution, as intuition expects. D^{term} divides by a scale estimated on the same pooled windows. Phantoms enter both numerator and denominator, and because $\rho < 1$ they depress the denominator proportionally more than the numerator at short horizons, inflating the ratio. The estimator's own scale is contaminated.

The consequence for practice is sharper than the mechanism. A user holding only a contaminated dataset cannot determine which regime they are in, because the diagnosis requires ρ, which requires knowing which items are phantoms. Statements of the form "contamination will attenuate my effect, so my estimate is conservative" — a very common defence in applied work — are not available. Attenuation is one of two possible directions and which one occurs depends on a quantity the analyst cannot measure.

5.5 A normalisation that returns 0.40 where the estimator returns 0.10

The pathology of §3.4 is not hypothetical; we found it in our own pipeline, and it accounts for a factor-of-four discrepancy that took two rounds to resolve. On an identical set of 335 events, the two aggregations of (2) give

$$\widetilde{D}(5\text{s}) = 0.3997 \qquad \text{versus} \qquad D(5\text{s}) = 0.1020.$$

Splitting by the magnitude of the reference return makes the source visible: on quiet windows ($|r_{\text{ref}}|$ below the sample median) $\widetilde{D}(5\text{s}) = 0.415$; on active windows, 0.097. The quiet windows, where by construction nothing happened and the post-event path is uncorrelated with the anchor sign, dominate the average because they carry the largest $1/|r_{\text{ref}}|$ weights.

Per-item normalisation is an appealing move — it is how one makes heterogeneous events comparable — and its failure is silent, because the estimator returns a plausible number that is stable across reruns. We report it because we suspect it is not rare: any statistic of the form "mean of per-item normalised response" with a signed, mean-zero denominator has this property.

5.6 Isolating the statistical channel

Finally we separate the mechanism from everything domain-specific. $M = 5000$ candidate windows per configuration, 10 seeds; a fraction π carry a delayed smooth jump ($\lambda \in \{2, 10, 60\}$ s, amplitude $U(3, 10) \times \sigma_{2\text{m}}$), the rest are pure random walk. Two detectors at *matched nominal strictness*: detector A thresholds $|\Delta p_{2\text{m}}|/\sigma_{1\text{s}}$ at 10; detector B

thresholds $|\Delta p_{2\mathrm{m}}|/(\sigma_{1\mathrm{s}}\sqrt{120})$ at 3. Both read as strict; their actual false-positive rates differ by two orders of magnitude (0.361 versus 0.0027). A self-check at $\pi = 1$ confirms the harness recovers the injected response (relative bias 0.039).

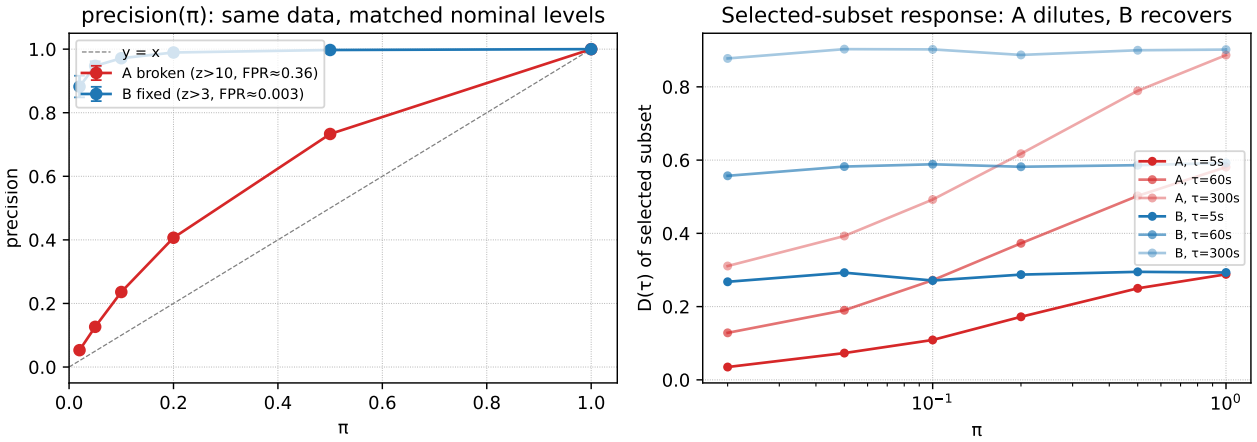

Figure 4: Two detectors at matched *nominal* strictness on identical data. **Left:** at $\pi = 0.02$, A attains precision 0.053 while B attains 0.882. **Right:** the response curve measured on each detector's selected subset, as a function of π: A's estimate drifts with the composition of its own selection; B's is stable.

At $\pi = 0.02$ the two equally strict detectors yield precisions of 0.053 and 0.882 on the same data. Nothing about the underlying events changed; only a time-scale normalisation inside the acceptance statistic did. This is the minimal mechanism behind Table 1: a detector-defined dataset is the detector's normalisation choices stamped onto data, and those choices are not recoverable from the output.

5.7 Scope, and a pre-registered claim we did not get to make

Before running the analysis we registered a threshold for an absolute claim: if the clean official index gave $D(5\mathrm{s}) \geq 0.30$ we would claim a fast price-discovery effect, and at ≤ 0.15 we would not. The measured value on 415 clean official windows is $\mathbf{0.068}$. We therefore make no such claim. We report this because the pre-registration is what makes the rest of the paper's numbers credible, and because the temptation to relax the threshold after seeing 0.068 was real and is exactly the behaviour the paper argues against.

6 Related work

Base rates and precision under imbalance. That posterior probability depends on prevalence is the base-rate fallacy (Bar-Hillel, 1980); its consequences for classifier evaluation are treated by Hand (2009), and Saito & Rehmsmeier (2015) show specifically that precision-based summaries are prevalence-dependent in a way that ROC summaries are not. Our contribution is not the mechanism but an end-to-end measurement of it in a deployed dataset with recoverable ground truth, plus the mixture and estimator-dependence results that follow.

Weak, distant, and self-supervision. Distant supervision (Mintz et al., 2009) and programmatic weak supervision (Ratner et al., 2017) construct labels by rule application; the label-model literature estimates and corrects labelling-function accuracies, but typically assumes the accuracies are pool-stable, which is precisely the assumption Table 1 violates. Pseudo-labelling suffers a related confirmation bias (Arazo et al., 2020). Positive-unlabelled learning (Elkan & Noto, 2008) formalises learning when negatives are unobserved and makes prevalence explicit; our setting is its evaluation-side dual.

Label noise. A large literature studies learning under label noise (Natarajan et al., 2013; Frénay & Verleysen, 2014). That work almost always models noise as a stochastic corruption applied to correct labels. The contamination here is not

of that form: phantoms are generated by the same rule that generates true positives, so the corruption is systematically correlated with the features and carries its own signal shape (§5.3). Noise-robust methods that assume class-conditional random flips do not address it.

Dataset quality and documentation. Northcutt et al. (2021) document pervasive label errors in benchmark test sets, and Gebru et al. (2021) propose documentation standards. Our protocol (§7) is in this tradition and specialises it to the detector-defined case, where the relevant fields are $(\pi, \mathrm{TPR}, \mathrm{FPR}, c, \rho)$ rather than provenance narrative.

Event studies. The applied setting is the event study (MacKinlay, 1997); the specific phenomenon our detector was built to find is announcement response (Andersen et al., 2003). We use this literature only as the source of a realistic detector and a realistic downstream estimand.

7 A reporting protocol for detector-defined datasets

1. **Observed indices first.** Use a published index wherever one exists; detection is a last resort, not a convenience.
2. **Validate in the deployment pool.** Precision must be measured, or derived via (1), on the pool where the detector is actually run. Cross-prevalence transfer of precision is not permitted. High-prevalence validation is near-zero evidence about FPR.
3. **Report five quantities.** π, TPR, FPR, c, ρ. Precision alone conceals which one moved, and ρ determines whether a given c is tolerable.
4. **Declare the time scale of every z-like statistic.** Any ratio of a scale-t quantity to a per-unit scale must be $\sqrt{t}$-normalised explicitly, and the reference window of every normalisation stated.
5. **Ship an analytic null with every estimator.** A closed-form null for the acceptance rule is what turns "our detector fires often" into a measured FPR. When the estimator is applied to a selected subset, the null must be the *selection-conditioned* null, not the unconditional one.
6. **Avoid per-item normalisation by a signed, mean-zero denominator.** Such estimators have no finite expectation (§3.4); use a pooled denominator or none.
7. **Run an oracle power check before comparing methods.** Compute the bound achieved by a method given perfect ground truth. If it is within seed noise of the baseline, the configuration cannot distinguish methods, and this is a property of the design rather than of the methods (App. B).

8 Limitations

1. **One detector family.** All measurements concern magnitude-threshold detectors. Equations (1) and (3) hold for any detector; our numbers do not transfer to change-point, likelihood-ratio, or learned detectors.
2. **One instrument, one asset class, one index family.** Prevalences and phantom rates are pool properties. The mechanism is general; the magnitudes are not.
3. **Sample coverage.** 415 clean official windows, effectively 2013 onward; 68 pre-2013 windows returned no data at the vendor's 1-second resolution. Cross-decade comparability is limited.
4. **The selection-conditioned null is incomplete.** It reproduces the shape of D_P but not its level (§5.3).
5. **Relative claims only.** Having abandoned the terminally normalised statistic for headline use, we support relative conclusions between datasets and estimators, not absolute statements about response speed.

Reproducibility statement

All event indices, phantom labels, screens, curves, and simulator code are released, together with a per-number traceability map, the full decision log for every pre-registered branch point, and an accident log (App. A). The price data

cannot be redistributed under the vendor's licence; we release the exact request manifests (symbol, schema, window, checksum) so that an identical dataset can be reconstructed.

References

T. Andersen, T. Bollerslev, F. Diebold, and C. Vega. Micro effects of macro announcements: real-time price discovery in foreign exchange. *American Economic Review*, 93(1):38–62, 2003.

E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *IJCNN*, 2020.

M. Bar-Hillel. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233, 1980.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. *KDD*, 2008.

B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE TNNLS*, 25(5):845–869, 2014.

T. Gebru, J. Morgenstern, B. Vecchione, J. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009.

A. C. MacKinlay. Event studies in economics and finance. *Journal of Economic Literature*, 35(1):13–39, 1997.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. *ACL-IJCNLP*, 2009.

N. Natarajan, I. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. *NeurIPS*, 2013.

C. Northcutt, A. Athalye, and J. Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *NeurIPS Datasets and Benchmarks*, 2021.

A. Ratner, S. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: rapid training data creation with weak supervision. *VLDB*, 11(3):269–282, 2017.

T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015.

A Six accidents in this paper's own pipeline

Each accident below occurred during construction of this paper's dataset, was resolved by a pre-registered action, and is documented in full in the released log. We present them as worked instances of the paper's mechanism. Each entry gives the error, how it was found, the blast radius, and the fix.

Accident 1: the unscaled z-score. The detector's statistic divides a 2-minute move by a *per-second* standard deviation, omitting the $\sqrt{120}$ scaling. Under a random-walk null this gives $\mathrm{FPR} = 2\big(1 - \Phi(10/\sqrt{120})\big) = 0.361$: the threshold fires on roughly one in three ordinary windows. Consequence: 487 phantom CPI detections. Table 1's 81.7% phantom rate *is this bug, measured*. Found by deriving the closed-form null and checking it against simulation (Rule 5), neither of which the original pipeline had. The bug is retained in §5.6 as detector A because it is realistic.

Accident 2: the hallucinated metadata correction. An automated pipeline step "corrected" the FOMC statement release time from 14:00 ET to 13:30 ET for all statements after a cutoff, citing a secondary report of a central-bank announcement that, on inspection of the institution's own press-release pages, does not exist. Both primary sources checked state release at 14:00 ET. Blast radius: 18 ground-truth timestamps shifted by 30 minutes, and one full response-curve computation silently rerun on the corrupted index before the error was caught. Fix: rollback; a hard rule that no metadata may be modified without a primary-source URL recorded in the manifest; per-statement verification against primary pages. This is the paper's thesis in miniature — a confident pipeline, an unmeasured false-positive rate on the task of "correcting metadata", and a ground truth that then validated itself.

Accident 3: positional indexing on a series with gaps. The first response-curve implementation located t_0 by row offset, assuming one row per second. The vendor emits a row only on quote update, so the offset drifted by up to 48 minutes in sparse periods and the signal averaged away. Fix: timestamp-based alignment throughout.

Accident 4: a cleaning filter that deleted the signal. A futures-roll filter flagged any 1-second return above 20 bp anywhere in the window. Of 86 official windows so flagged, 68 had their largest jump within ± 5 s of t_0: the filter was removing announcement reactions as artefacts. Fix: the roll screen is evaluated strictly on the pre-window, which cannot contain the reaction. *Lesson, and a general one:* any cleaning rule whose support intersects the event window can become selective deletion of the target signal — structurally identical to the contamination this paper studies, with the sign flipped.

Accident 5: the same indexing bug, in a different file. After Accident 3 was fixed in the estimator, the identical row-offset pattern reappeared in the screening code's pre-window slice. In high-activity sessions the "pre-window" slice crossed t_0 and the largest announcement reactions were deleted as rolls — for instance the November 2022 CPI release, whose true pre-window maximum was 3.3 bp and whose post-event move was 126 bp. The months with the strongest signal were deleted at the highest rate. Fix: timestamp masks, a repository-wide audit for row-versus-time indexing, and a regression test. Clean official windows went from 343 to 415.

Accident 6: a factor-of-four estimator artefact. Two implementations of the response curve disagreed by $4\times$ on identical events. The cause was the per-item normalisation of §5.5. Resolution is reported in the main text rather than only here because we consider the estimator pathology a finding rather than a mishap.

B Oracle power checks and a retracted experiment

For any timing-sensitive evaluation, the *oracle* is the same downstream method given the true timestamps. The oracle-minus-baseline gap upper-bounds the total effect any alignment or denoising method can produce in that configuration. An earlier version of this project evaluated an alignment module in a configuration whose oracle gap was 0.024 CRPS against a comparable seed-noise floor: no method comparison there could have detected anything, regardless of method quality. The experiment was retracted before producing conclusions. Rule 7 operationalises the check as a pre-registration step. The retracted results are released for completeness and should not be cited.

C The revision-only negative control

Same-month double releases — where a scheduled release carries only a revision to previously published figures and no new reference-period data — provide events with the exact metadata structure of real releases but, by construction, no new information. A calendar rule identifies 17 such windows (13 CPI, 4 payrolls). Only 3 have usable price coverage in our sample. We report this because the control was pre-registered, and we do *not* interpret it: $n = 3$ is insufficient, and the curve shown in Figure 3 is included only for completeness. Constructing this control properly requires price coverage we could not obtain within the project's data budget, and we flag it as the single cheapest improvement available to a follow-up.

D The terminally normalised statistic

For completeness we give the legacy 2-hour-anchor convention on the screened sample. At $\tau = 5$ s: official 0.068, detected 0.082; at 300 s: 0.271 and 0.317; random-walk baseline $\tau/7200$. This is the statistic underlying the original dataset's published curve. Its ordering (detected above official at every horizon) reverses under $S(\tau)$, which is the demonstration of §5.4. Per-category values at $\tau \in \{5, 30, 300, 3600\}$.

E Detector specification and screens

The detector scans the front-month continuous ES series at 1-second best-bid-offer resolution and accepts a window when $|\Delta p_{2\mathrm{m}}|$ exceeds ten per-second standard deviations estimated on the 5 minutes preceding t_0 (Accident 1's unscaled form). Windows are $t_0 \pm 2\,\mathrm{h}$ on bid-ask mid. Screens: (i) pre-window roll screen, any 1-second log return above 20 bp strictly before $t_0 - 1\,\mathrm{s}$, evaluated on timestamp masks; (ii) degraded-quote screen, windows with more than 5% missing bid or ask. Screens are applied identically to every label set before any comparison.

Use of AI tools

This section is required by the ICLR 2027 AI policy and does not count toward the page limit. We state the extent of AI involvement in full.

We used generative AI tools (a large language model with agentic tool use, and a coding agent built on it) for the following tasks in the policy's *required-disclosure* list: developing the conceptual framework of this paper; formulating and refining the hypotheses; formulating the mathematical claims in §3, including the random-walk baseline, the mixture identity (3), and Proposition 1; supplying the argument used in the proof of Proposition 1; designing the experimental protocol and the pre-registered decision rules; implementing all data acquisition, screening, estimation, and simulation code; cleaning and reformatting the event index; and interpreting results. We also used them for tasks in the *recommended-disclosure* list: producing the figures, drafting and editing the text, identifying and formatting references, surveying related work, and proposing the title and section structure.

Three points bearing on how this disclosure should be read. First, the project was structured around pre-registered decision rules with numeric thresholds fixed before the corresponding measurement, together with a rule that thresholds could not be revised after seeing results. Section 5.7 reports a threshold that was not met and a claim consequently not made; Limitation 4 reports a second arm terminated at a pre-registered health check. These are the mechanism by which AI-assisted work was constrained, and they are visible in the released decision log rather than asserted here.

Second, the AI tooling introduced errors, and we document them rather than only the successes. Appendix A lists six. Accident 2 is the most instructive: an automated step fabricated a citation to a central-bank announcement that does not exist and used it to "correct" 18 ground-truth timestamps, after which a full analysis was silently rerun on the corrupted index. It was caught by manual inspection of primary sources. We report it in the main line of the paper because it is a worked instance of the paper's own thesis, and because it is the reason our project rule now forbids any metadata modification without a primary-source URL recorded in the manifest.

Third, and consequently: verification in this project was manual and adversarial, not incidental. Primary sources were checked by hand for every metadata correction. Estimator discrepancies were resolved by deriving closed forms and checking them against simulation rather than by accepting whichever number looked more plausible. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with AI assistance.