ROBUST CONFORMAL ANOMALY DETECTION UNDER CALIBRATION POISONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction is routinely used to wrap an anomaly detector in a distribution-free coverage guarantee. That guarantee is about normal samples only, and we show it can be satisfied by a detector that has stopped working. When an attacker places a small fraction of high-scoring points into the calibration set, the conformal threshold moves past the range of plausible normal scores, so coverage on normal data is met or exceeded while recall on true anomalies falls to near zero. We give a closed-form account: the operating level of standard split conformal under upper-tail contamination is $\min\{(1-\alpha)/(1-\epsilon), 1\}$, so at exactly $\epsilon = \alpha$ the detector stops raising false alarms and keeps only whatever power comes from anomalies scoring beyond the clean support, a collapse point predictable before any experiment is run. We then analyze a trimmed-quantile calibration rule with the level correction $\alpha' = \alpha(1+2\epsilon)$. Writing its threshold as an order statistic of the full calibration sample reduces the analysis to two short concentration arguments and yields a two-sided bracket on the operating level whose upward drift is $O(\alpha\epsilon^2)$ rather than $O(\epsilon)$, a power bound, and an expression for attainable coverage that shows where a symmetric trim can be driven and motivates a one-sided variant needing no contamination parameter. Across 5 tabular benchmarks, 5 detectors, 3 attacker models, and frozen CLIP and MiniLM embeddings, standard conformal retains 12 to 19% of its clean power at $\epsilon = 0.10$ and 4 to 7% at $\epsilon = 0.30$, and 0.1% on an autoencoder, while the trimmed rule stays above 97% everywhere.

1 INTRODUCTION

Anomaly detection sits underneath fraud screening, intrusion detection, industrial monitoring, and clinical triage (Chandola et al., 2009; Pang et al., 2021). Detectors emit uncalibrated scores, so a common engineering step is to wrap a trained detector in split conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2018) and turn scores into decisions with a distribution-free guarantee: at most an α fraction of normal inputs will be flagged. The guarantee holds for any score function and any data distribution, and it can be checked in production by monitoring the false alarm rate. That checkability is where the trouble starts, because conformal coverage constrains the system on normal inputs only. It says nothing about anomalies, and a detector that flags nothing at all satisfies it perfectly. In ordinary operation the two quantities move together, so the gap is easy to overlook; under contamination of the calibration set they come apart, and in the direction that hides the failure.

The mechanism takes one sentence to state. Split conformal places its threshold at an upper order statistic of the calibration scores, so any calibration point scoring higher than a genuine normal point pushes that order statistic up. An attacker who gets a few high-scoring points into the calibration stream raises the threshold, and once the threshold clears the support of the normal score distribution the false alarm rate goes to 0 and coverage goes to 1. Whoever watches coverage sees a number better than the target; whoever watches recall, if anyone does, sees the detector stop firing. Calibration data is a realistic thing to attack: systems that recalibrate on recent traffic labeled normal, fraud pipelines calibrated on transactions nobody flagged, and moderation pipelines calibrated on content that survived review all admit material the adversary influenced, with no access to weights, code, or the training store (Biggio et al., 2012; Jagielski et al., 2018).

Robust statistics supplies an obvious repair, replacing the empirical quantile with a trimmed one (Huber, 1964; Rousseeuw, 1984), and what we take seriously is what that substitution buys when

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.

contamination is chosen adversarially, reaches 30%, and no labels are available, a regime outside the scope of existing analysis of conformal outlier detection under a contaminated reference set (Bashari et al., 2025). Our starting point removes most of the technical difficulty: the trimmed quantile is not a new estimator needing its own empirical process theory, but an ordinary order statistic of the full, independent and identically distributed calibration sample evaluated at a shifted level. Concentration then applies to the full sample rather than to a rank-selected subsample, and results can be stated in terms of the operating level $F_P(\theta)$ rather than the threshold θ, which removes the need to invert a distribution function and any assumption about the density near the target quantile.

Contributions.

- We characterize the failure mode quantitatively. Under upper-tail calibration contamination the operating level of standard split conformal is $\min\{(1-\alpha)/(1-\epsilon), 1\}$, so its type-I error reaches 0 at exactly $\epsilon = \alpha$, taking detection power with it for any anomaly distribution supported below the inflated threshold (Proposition 9). The collapse point is a prediction, not a fitted observation.
- We analyze the trimmed-quantile rule with correction $\alpha' = \alpha(1 + 2\epsilon)$ in level space. Lemma 3 identifies its threshold as the order statistic at level $p^\star = 1 - \epsilon - \alpha(1 - 4\epsilon^2) + O(1/m)$, giving a two-sided bracket on $F_P(\theta)$ whose only regularity requirement is continuity of the score distributions (Theorem 5) and a power bound under a checkable separation condition (Corollary 7). The upward drift is $4\alpha\epsilon^2$, second order in the contamination rate and independent of the contaminating distribution, though the accompanying concentration term is what dominates at the sample sizes we run (Remark 6).
- We give attainable coverage in closed form, as $(p^\star - c/m)/(1-\epsilon)$ in the number c of contaminated points below the threshold (Proposition 8). Being linear in placement, it identifies the configuration in which a symmetric trim does worst, which motivates RCAD-U, a one-sided rule that estimates tail excess from a clean reference, takes no contamination parameter, and has clean-sample level pinned at $1 - \alpha$ for every placement.
- We evaluate on 5 tabular benchmarks, 5 detectors spanning isolation, density, kernel, distributional and reconstruction-based scoring, 3 attacker knowledge models, 4 contamination placements, and frozen CLIP and MiniLM embeddings, and report where the defense helps, where it is unnecessary, and what it costs.

The rest of the paper is organized as follows. Section 2 reviews related work, Section 3 fixes notation and the threat model, states the two calibration rules and develops the theory, Section 4 describes the experimental setup, Section 5 reports the experiments, Section 6 draws out what they imply for deployed conformal detectors, and Section 7 concludes. Proofs and additional experiments are in the appendix.

2 Literature Review

Conformal prediction and violated exchangeability. Split conformal prediction converts any score function into a set-valued predictor with finite-sample marginal coverage, provided calibration and test points are exchangeable (Papadopoulos et al., 2002; Vovk et al., 2005; Shafer and Vovk, 2008; Lei et al., 2018; Angelopoulos and Bates, 2023), with extensions to conditional coverage through quantile regression (Romano et al., 2019) and to leave-one-out constructions (Barber et al., 2021). Exchangeability is the assumption that breaks first in practice, and two lines of work address that: reweighting under known covariate shift (Tibshirani et al., 2019), and coverage bounds that degrade with a distance between the calibration and test distributions (Barber et al., 2023). Both suppose the departure is a property of the environment, whereas calibration poisoning is chosen, and chosen with knowledge of what the quantile estimator will do with it. Closest in spirit is adversarially robust conformal prediction (Gendler et al., 2022), which perturbs inputs at test time and leaves calibration intact.

Conformal outlier detection under contamination. One-class conformal procedures calibrate on normal data only and report a p-value per test point (Laxhammar and Falkman, 2015; Bates et al., 2023; Guan and Tibshirani, 2022); what they inherit is type-I error control on normal inputs, with power left to the base score function. Most relevant here, Bashari et al. (2025) study conformal outlier

detection with a contaminated reference set, show contamination leaves type-I error conservative while costing power, and recover power with an annotation budget: their Label-Trim procedure spends labels on the most suspicious calibration points and removes the ones an oracle confirms, against a Standard baseline and a Naive-Trim baseline that drops top scores blindly. We keep their baselines and ask what happens when contamination is adversarially placed, reaches 30%, and no labeling budget exists.

Robust estimation, detectors, and poisoning. Trimmed and Winsorized estimators have a well understood breakdown point and a long history as the standard response to contaminated samples (Huber, 1964; Rousseeuw, 1984; Hampel et al., 1986; Huber and Ronchetti, 2009), so our contribution is not the estimator but the observation that inside a conformal procedure the trimmed quantile reads as a full-sample order statistic at a shifted level, which makes an exact level calculation possible. We treat the base detector as a black box and evaluate across families: isolation-based (Liu et al., 2008), density-based (Breunig et al., 2000), one-class kernel methods (Schölkopf et al., 2001), empirical distribution scores (Li et al., 2023), and reconstruction-based neural detectors (Ruff et al., 2018), following the benchmarking practice of Han et al. (2022) and Pimentel et al. (2014). Poisoning of training data is well established for supervised models (Biggio et al., 2012; Steinhardt et al., 2017; Jagielski et al., 2018); our setting differs in target, since the fitting set is untouched and only the calibration set is contaminated, which isolates the effect on the decision threshold from any effect on the score function.

3 METHODOLOGY

3.1 SETUP AND THREAT MODEL

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space and P the distribution of normal data. A base detector is a score function $A : \mathcal{X} \to \mathbb{R}$ with larger values more anomalous. Write F_P for the distribution function of $A(X)$ when $X \sim P$ and F_Q for its counterpart under the contaminated calibration law Q defined below. We assume throughout that both are continuous, so that calibration scores are almost surely distinct; continuity is the only regularity condition we use, and ties, where a detector produces them, are broken at random. Training data is split into a fitting set $\mathcal{D}_{\text{fit}}$, used to fit A, and a calibration set $\mathcal{D}_{\text{cal}} = \{X_1, \dots, X_m\}$, used to set a threshold θ. A test point is flagged when $A(x) > \theta$, and two quantities describe the detector,

$$\text{Cov}(\theta) = \mathbb{P}_{X\sim P}[A(X) \le \theta] = F_P(\theta), \qquad \text{Pow}(\theta) = \mathbb{P}_{X\sim P_{\text{anom}}}[A(X) > \theta]\,. \tag{1}$$

Coverage is 1 minus the type-I error, and we call $F_P(\theta)$ the *operating level*; power is recall on test-time anomalies. A detector with $\theta = +\infty$ attains coverage 1 and power 0, so coverage alone cannot certify that a detector works. Standard split conformal sets $\theta_{\text{std}} = A_{(\lceil (1-\alpha)(m+1) \rceil)}$ from the sorted calibration scores $A_{(1)} \le \cdots \le A_{(m)}$ and reports the conformal p-value $p(x) = (|\{i : A_i \ge A(x)\}| + 1)/(m+1)$.

We use Huber's contamination model (Huber, 1964): the calibration sample is drawn independently from

$$Q = (1-\epsilon)P + \epsilon P_{\text{cont}}, \qquad \epsilon \in [0, 1/2), \tag{2}$$

with P_{cont} arbitrary. The fitting set is never modified, so the score function is the clean one and every effect we measure is attributable to the threshold. Regimes differ in how P_{cont} is chosen: *in-distribution*, where P_{cont} is the natural anomaly distribution and models incomplete cleaning; *heavy-tail*, where P_{cont} places mass far outside the support of P; and *adversarial*, where the attacker scores a candidate pool and submits those with the largest nonconformity. We also vary the *placement* of contaminated mass, which turns out to matter most for a symmetric rule. The attacker needs no access to weights or code, only the ability to place points into the stream the calibration set is drawn from and enough knowledge of the score distribution to know which points score high.

3.2 TWO CALIBRATION RULES

Definition 1 (RCAD, known contamination budget)**.** Given α and an assumed budget ϵ, set $k = \lfloor \epsilon m \rfloor$ and $\alpha' = \alpha(1 + 2\epsilon)$. The threshold is the order statistic

$$\theta = A_{(j)}, \qquad j = \lceil (1-\alpha')(m-2k) \rceil + k. \tag{3}$$

Algorithm 1 Calibration and prediction

Require: training data, calibration fraction γ, level α; a budget ϵ (RCAD) or a reference sample $\mathcal{D}_{\mathrm{ref}}$ (RCAD-U)

1: Split training data into $\mathcal{D}_{\mathrm{fit}}$ and $\mathcal{D}_{\mathrm{cal}}$ with $|\mathcal{D}_{\mathrm{cal}}| = m$, fit A on $\mathcal{D}_{\mathrm{fit}}$, and sort $A_{(1)} \le \cdots \le A_{(m)}$
2: **RCAD:** $k \leftarrow \lfloor \epsilon m \rfloor$, $\alpha' \leftarrow \alpha(1+2\epsilon)$, $j \leftarrow \lceil (1-\alpha')(m-2k) \rceil + k$
3: **RCAD-U:** score $\mathcal{D}_{\mathrm{ref}}$, form $\widehat{\epsilon}_{\mathrm{up}}, \widehat{\epsilon}_{\mathrm{low}}$ by Eq. (4), $j \leftarrow j_U$ by Eq. (5)
4: $\theta \leftarrow A_{(j)}$; **predict:** flag x as anomalous if $A(x) > \theta$

Equation (3) is the $(1-\alpha')$ empirical quantile of the sample with k points removed from each tail, and the correction $\alpha' > \alpha$ lowers the target level to compensate for the upward bias contamination induces; the estimator tolerates k corruptions per tail (Proposition 12). The analysis below sets the assumed budget equal to the true contamination rate and writes ϵ for both; Section 5.4 measures what misspecification costs. The exact coverage expression in Proposition 8 shows a symmetric trim is the wrong shape for a one-sided threshold: trimming from the bottom does nothing against upper-tail contamination, while trimming from the top discards legitimate high scores, which is what an adversary placing mass at the bottom exploits. The next rule fixes this and removes the ϵ input.

Definition 2 (RCAD-U, reference driven)**.** Let $\mathcal{D}_{\mathrm{ref}}$ be a held-out reference sample believed clean, with empirical distribution function $\widehat{F}_{\mathrm{ref}}$, and let $\widehat{F}_m$ be that of the calibration scores. Estimate the one-sided tail excesses

$$\widehat{\epsilon}_{\mathrm{up}} = \sup_t \left[\widehat{F}_{\mathrm{ref}}(t) - \widehat{F}_m(t) \right]_+, \qquad \widehat{\epsilon}_{\mathrm{low}} = \sup_t \left[\widehat{F}_m(t) - \widehat{F}_{\mathrm{ref}}(t) \right]_+, \tag{4}$$

set $k_{\mathrm{up}} = \lfloor \widehat{\epsilon}_{\mathrm{up}} m \rfloor$, $k_{\mathrm{low}} = \lfloor \widehat{\epsilon}_{\mathrm{low}} m \rfloor$, and take

$$\theta_U = A_{(j_U)}, \qquad j_U = \lceil (1-\alpha)(m - k_{\mathrm{up}} - k_{\mathrm{low}}) \rceil + k_{\mathrm{low}}. \tag{5}$$

The rule trims k_{up} points from the top and, rather than trimming the bottom, shifts the target index up by k_{low} so the retained lower mass is skipped. Its level within the retained block is then $(j_U - k_{\mathrm{low}})/(m - k_{\mathrm{up}} - k_{\mathrm{low}}) = 1 - \alpha$ up to the rounding of the ceiling, whatever the placement, which is the property the symmetric rule lacks. That level transfers to the clean distribution to the extent that the trimmed points are the contaminated ones, so the rule depends on the tail-mass estimates rather than on a supplied budget. Algorithm 1 states both rules; calibration costs one sort, $O(m \log m)$, on top of the detector fit, and prediction is a single comparison.

3.3 Analysis

Proofs are in Appendix A.

Lemma 3 (Order statistic identity)**.** *The RCAD threshold of Definition 1 is the j-th order statistic of the full calibration sample, and its level in that sample satisfies*

$$\frac{j}{m} = p^\star + O(1/m), \qquad p^\star = 1 - \epsilon - \alpha\left(1 - 4\epsilon^2\right), \tag{6}$$

with $|j/m - p^\star| \le 2/m$. Moreover $\partial p^\star / \partial \epsilon = -1 + 8\alpha\epsilon < 0$ for all $\epsilon < 1/2$ and $\alpha < 1/4$.

The identity moves the analysis off the trimmed subsample, which matters: a rank-selected subsample is neither independent nor identically distributed, so the Dvoretzky–Kiefer–Wolfowitz inequality does not apply to it (Dvoretzky et al., 1956; Massart, 1990), while it does apply to the full sample.

Lemma 4 (Uniform deviation under contamination)**.** *Let $\widehat{F}_m$ be the empirical distribution function of m independent draws from Q in Eq.* (2)*. With probability at least $1-\delta$, $\sup_t |\widehat{F}_m(t) - F_P(t)| \le \epsilon + \sqrt{\log(2/\delta)/2m}$.*

The bias term is elementary: $F_Q - F_P = \epsilon(F_{\mathrm{cont}} - F_P)$ and both distribution functions take values in $[0,1]$, so the gap is at most ϵ pointwise with no assumption on P_{cont}. Being uniform in t, the bound may be evaluated at the data-dependent point θ, which is what makes the next result possible without inverting F_P.

Theorem 5 (Two-sided bracket on the operating level)**.** *Let $s_{m,\delta} = \sqrt{\log(2/\delta)/2m} + 2/m$. With probability at least $1-\delta$, the RCAD threshold satisfies*

$$1 - \alpha - 2\epsilon - s_{m,\delta} \;\le\; F_P(\theta) \;\le\; 1 - \alpha + 4\alpha\epsilon^2 + s_{m,\delta}. \tag{7}$$

The two sides are not symmetric, and that asymmetry is the design goal: the level can drift down by $O(\epsilon)$, which costs coverage, but up by only $4\alpha\epsilon^2$, which is what protects power. That is what the correction $\alpha' = \alpha(1+2\epsilon)$ is for, since the trimming shift of $-\epsilon$ in Eq. (6) offsets the worst-case contamination bias of $+\epsilon$ in Lemma 4, leaving no dependence on P_{cont}.

Remark 6 (Which term dominates)**.** Theorem 5 describes the structure of the drift, not a tight finite-sample certificate. At $\delta = 0.05$ the concentration term is $s_{m,\delta} = 0.106$ at $m = 200$ and 0.025 at $m = 3200$, so it dwarfs $4\alpha\epsilon^2 = 0.004$ and the upper bound exceeds 1, hence says nothing, for $m < 240$; Corollary 7 is informative on the same schedule. What the theorem establishes is that contamination enters at second order while the sampling term decays at the usual rate, so the two separate, and Proposition 8 carries the finite-sample content the experiments test.

Corollary 7 (Power retention)**.** *Write $\gamma = 4\alpha\epsilon^2 + s_{m,\delta}$ and let π_0 be the power at the clean threshold $F_P^{-1}(1-\alpha)$. On the event of Theorem 5,*

$$\text{Pow}(\theta) \;\ge\; \pi_0 - \mathbb{P}[F_P(A(X_{\text{anom}})) \in (1-\alpha,\, 1-\alpha+\gamma]]\,. \tag{8}$$

If the anomaly distribution places no mass in that probability integral transform band, power is at least π_0.

The separation condition can be checked rather than assumed, by fitting F_P on held-out clean scores and inspecting the transformed anomaly scores on a band of width γ above $1-\alpha$.

Proposition 8 (Attainable coverage)**.** *Condition on a calibration set containing m_0 clean points and c contaminated points at or below θ. The fraction of clean calibration points at or below θ is then exactly $(j-c)/m_0$, and since $m_0/m \to 1-\epsilon$,*

$$\text{Cov} = \frac{p^\star - c/m}{1-\epsilon} + O(1/m) + O_p(m^{-1/2}), \tag{9}$$

linear in c and therefore in contamination placement, with endpoints $p^\star/(1-\epsilon)$ when all contamination sits above the threshold and $(p^\star-\epsilon)/(1-\epsilon)$ when all of it sits below.

At $m = 200$ and $\alpha = \epsilon = 0.10$ the interval is $[0.7822, 0.8933]$ against a target of 0.900, with measured values 0.8917 under upper-tail contamination and 0.7766 under a bottom-placed adversary, both inside the $O_p(m^{-1/2}) \approx 0.024$ fluctuation of the clean count at this sample size. The gap between the endpoints is why RCAD-U exists, and also why Eq. (7) keeps two terms: a single $C\sqrt{\epsilon/m}$ expression cannot absorb them at any constant, since the deficit it must cover is $\Theta(\epsilon)$ and does not shrink with m, so the required C grows like $\sqrt{m}$, already 5.27 at $m = 200$ and 21.1 at $m = 3200$.

Proposition 9 (Saturation of standard conformal)**.** *Under contamination placed entirely above the clean support, the operating level of standard split conformal converges to*

$$F_P(\theta_{\text{std}}) \;\longrightarrow\; \min\left\{\frac{1-\alpha}{1-\epsilon},\, 1\right\}, \tag{10}$$

which reaches 1 at exactly $\epsilon = \alpha$. At and beyond that point the type-I error is 0, and the power is 0 for any anomaly distribution supported below θ_{std}.

Proposition 9 is the quantitative form of the failure mode: at $\alpha = 0.10$ a conformal detector stops raising false alarms once 10 % of its calibration stream is upper-tail contamination, and whatever detection power survives comes only from anomalies scoring above a threshold that now sits beyond the clean support. The statement is fixed before any experiment is run, and Section 5.1 tests it.

Remark 10 (Conformal p-values after trimming)**.** The statements above concern the decision rule $A(x) > \theta$. A p-value computed against the untrimmed calibration set does not inherit them, so we treat such p-values as heuristic scores, not calibrated quantities.

4 EXPERIMENTAL SETUP

Data and detectors. Five tabular benchmarks: breast cancer Wisconsin (569 samples, 30 features, 37.0% anomalies), mammography (11,183 samples, 6 features, 2.3 %), satellite (5,100 samples, 36 features, 1.5%), a synthetic moderate-dimensional problem at a 10% anomaly rate, and digits with the digit 8 as the anomaly class, all from OpenML with multi-class targets mapped to majority against rest (Han et al., 2022). For representation-space experiments we use frozen CLIP image embeddings (Radford et al., 2021) on CIFAR-10, with classes 1 and 9 each taken as the anomaly class in turn, and frozen MiniLM sentence embeddings (Reimers and Gurevych, 2019) on 20 Newsgroups with class 7. Detectors are Isolation Forest (Liu et al., 2008), Local Outlier Factor (Breunig et al., 2000), one-class SVM (Schölkopf et al., 2001), ECOD (Li et al., 2023), and a multilayer perceptron autoencoder scored by reconstruction error, with k nearest neighbor and Mahalanobis distance on embeddings; implementations follow Pedregosa et al. (2011), with scores negated where needed so larger always means more anomalous.

Protocol. For each dataset and seed we split into train and test, split the clean training normals into $\mathcal{D}_{\text{fit}}$ and a clean calibration set of size $m = 200$ ($m = 600$ for the clean-coverage study), and replace $\lfloor \epsilon m \rfloor$ calibration points with contaminated points from the regime under test, leaving the fitting set untouched. Unless stated otherwise $\alpha = 0.10$, $\gamma = 0.30$, and results are means over 8 seeds.

Attackers. Three knowledge models, detailed in Appendix E. *White box* scores a candidate pool under the defender's own detector and submits the top k by nonconformity; *partial knowledge* fits its own Isolation Forest on a disjoint public sample carved out of the defender's fitting set and never observes the defender's detector; and *matched capacity* gives a white-box attacker that same reduced fitting set, so the last two arms differ in exactly one variable. We also sweep contamination placement over 4 positions from the top of the score distribution to the bottom.

Baselines and metrics. Standard split conformal, the Standard flavor of Bashari et al. (2025); Naive-Trim, which drops the top k scores with no level correction; and Label-Trim (Bashari et al., 2025), which is given oracle labels on the calibration set and is therefore an upper reference point rather than a peer. We report type-I error and power at the decision threshold as in Eq. (1), and *power retention*, the ratio of power at contamination level ϵ to the same configuration's power at $\epsilon = 0$, so detectors of different strength can be compared without a shared baseline. Where a ratio would have a very small denominator, we report absolute power instead.

5 RESULTS

5.1 THE COLLAPSE, AND THE PREDICTED COLLAPSE POINT

Figure 1 shows the white-box arm on Isolation Forest, pooled over the 5 tabular benchmarks and over contamination placements. Absolute values differ between this pooled view, the detector sweep below and the fixed upper-tail placement of Section 5.4, so we compare within an experiment and report retention across them. Standard conformal loses power from 0.4661 at $\epsilon = 0$ to 0.0791 at $\epsilon = 0.10$ and 0.0282 at $\epsilon = 0.30$, a retention of 17.0 and 6.0%, while its type-I error falls from 0.0924 to 0.0017, roughly 54 times below its own clean rate: the detector is not trading errors, it has stopped firing. RCAD holds type-I error between 0.0926 and 0.1047 against a nominal 0.100, and power between 0.4609 and 0.4805.

Per dataset at $\alpha = 0.10$, standard conformal loses power from 0.83 to 0.06 on breast cancer, from 0.66 to 0.01 on mammography, and from 0.84 to 0.08 on satellite between $\epsilon = 0$ and $\epsilon = 0.30$, with satellite already at 0.27 by $\epsilon = 0.10$, while its coverage reaches 1.00 on 4 of the 5 benchmarks at $\epsilon \geq 0.10$; RCAD keeps coverage within 0.02 of target and power within 0.05 of the clean baseline throughout. Equation (10) fixes the collapse point at $\epsilon = \alpha$ with no free parameter, and standard conformal's measured type-I error is 0.0924 at $\epsilon = 0$ and 0.0036 at $\epsilon = 0.10$, so it stops raising false alarms where the expression says it should.

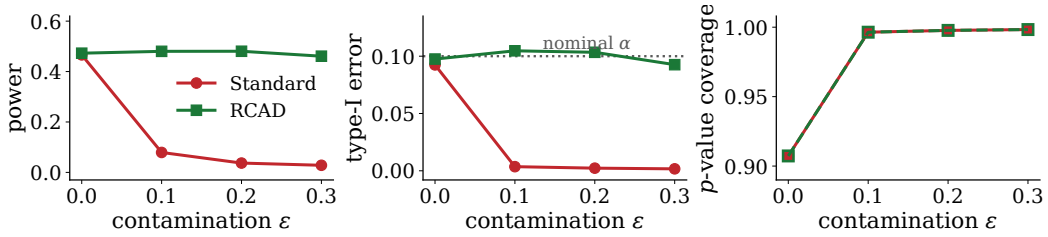

Figure 1: Standard conformal (circles) against RCAD (squares) under white-box calibration poisoning, Isolation Forest, $\alpha = 0.10$, pooled over 5 tabular benchmarks and 8 seeds. Left: power. Center: type-I error, nominal level dotted. Right: coverage from conformal p-values against the untrimmed calibration set, which reads the same to 4 decimals for both methods at every contamination level.

5.2 The effect is not specific to one detector family

Standard conformal retains 12.3 to 18.5% of its clean power at $\epsilon = 0.10$ and 4.1 to 7.2% at $\epsilon = 0.30$ across Isolation Forest, Local Outlier Factor, and one-class SVM, while RCAD retains 100.5 to 102.1% and 96.9 to 97.5% (Figure 2, left). ECOD, whose scores come from marginal empirical distribution functions rather than distances or partitions, shows the same pattern, so the mechanism is not an artifact of tree or kernel scores. The autoencoder is the sharpest case and also the strongest detector we measured, at clean power 0.5837 in this sweep against 0.3800 for Isolation Forest: standard conformal power on it falls to 0.0006 at $\epsilon = 0.10$, a factor of 973, while RCAD holds power between 0.59 and 0.61 with coverage between 0.892 and 0.903. The failure mode is therefore not confined to weak base detectors, which is what one would expect if the collapse were driven by poor score separation rather than by the threshold.

How much the attacker needs to know. At $\epsilon = \alpha = 0.10$, power retention across the white-box, partial knowledge, and matched capacity arms runs 0.170, 0.146, and 0.191 for standard conformal against 1.016, 1.030, and 1.038 for RCAD. Comparing the two arms that differ only in attacker knowledge, the surrogate attacker destroys 85.4% of standard conformal's power and the white-box attacker with the same detector capacity destroys 80.9%, so seeing the deployed detector does not help the attack here and is worth -4.5 percentage points, a gap of the size these 40 trials can resolve only weakly. The attack needs the shape of the score distribution, not the identity of the model. Bootstrapping one-sided 90% bounds over the 40 dataset by seed trials, RCAD's retention floor is at least 0.917 in every arm at every contamination level, against at most 0.032 for standard conformal.

5.3 Coverage diagnostics do not see any of this

Coverage computed from conformal p-values against the untrimmed calibration set reads 0.9073, 0.9964, 0.9977, and 0.9983 as ϵ goes from 0 to 0.30, identically to 4 decimals for both methods, since both use the same calibration set (Figure 1, right). At $\epsilon = 0.30$ it reports 0.9983 for a detector holding type-I error at 0.0926 with power 0.4609 and for one at type-I error 0.0017 with power 0.0282. The same insensitivity appears with no attacker present: on clean data, sweeping the assumed ϵ from 0 to 0.30 moves the type-I error at the trimmed threshold from 0.0975 to 0.3659 while the diagnostic reads 0.9073 throughout, and it cannot move, since the calibration set is identical across the sweep. This is why we report type-I error at the decision threshold everywhere.

5.4 Misspecified ϵ, and doing without it

Figure 3 in Appendix C sweeps assumed against true ϵ independently. On the diagonal, type-I error lands on 0.0999, 0.1057, and 0.1037 against a nominal 0.100 at 3 true contamination rates with nothing tuned, so the parameter has a correct value rather than being a knob that trades one rate against another. Below the truth, type-I error and power fall together, at a true $\epsilon = 0.20$ giving 0.0000 and 0.0624 against 0.1037 and 0.7189 on the diagonal, a direction conservative for validity since $\partial p^\star / \partial \epsilon < 0$; above the truth, type-I error inflates to 3.3 times nominal at a true 0.05 with an assumed 0.30.

RCAD-U removes the parameter and separates the two attack directions cleanly: under upper-tail contamination $\widehat{\epsilon}_{\mathrm{up}}$ reads 0.001, 0.053, and 0.244 at true ϵ of 0, 0.10, and 0.30 while $\widehat{\epsilon}_{\mathrm{low}}$ stays near

zero, and the two swap under bottom-placed contamination. On clean data its coverage is 0.9027, 0.9008, and 0.9024 on 3 detectors against a nominal 0.900 with power matching standard conformal to within 0.001, so it costs nothing absent an attack. Under attack it gives up power relative to a correctly specified RCAD: at $\epsilon = 0.10$ on Isolation Forest under fixed upper-tail contamination it reaches coverage 0.9541 and power 0.2969 against 0.8917 and 0.3858 for RCAD, so 77% of RCAD's power and 5.7 times standard conformal's 0.0519. What it depends on is the reference sample: as n_{ref} grows from 40 to 320 its power rises monotonically from 0.267 to 0.526, while $\widehat{\epsilon}_{\mathrm{up}}$ reads 0.043, 0.031, 0.047, 0.052 against a true rate near 0.05, noisy at small n_{ref} rather than monotone (Table 4).

Contamination placement. Coverage is linear in the number of contaminated points below the threshold by Proposition 8, and over the 15 position-by-rate cells given by 5 placement positions and 3 contamination rates, each averaged across the 5 benchmarks, the largest deviation from Eq. (9) is 0.0087. Over 2,400 configurations spanning 3 detectors, 5 datasets, 8 seeds, 4 placements and ϵ up to 0.30, the worst coverage reached by a correctly specified symmetric RCAD is 0.5002 on Isolation Forest and 0.5307 on the autoencoder, produced by bottom-placed rather than upper-tail contamination, against 0.8572 and 0.8608 for standard conformal. RCAD-U's worst cases are 0.9027 and 0.9024, which are its clean-data values, since its coverage only rises under attack; across all 3 detectors it never falls below 0.9008.

5.5 Frozen foundation model embeddings

On CLIP image and MiniLM sentence embeddings at a contamination rate of 0.20, the ordering found on tabular data holds in all 6 cells, with standard conformal retaining 53 to 59% of its clean power across the 4 CIFAR-10 configurations against 102 to 115% for RCAD. It retains far more power here than on tabular data, and Eq. (9) says why: complete shutdown requires contamination above the clean support, and in embedding space the attack does not get there, with only 0.1 to 24% of injected points scoring above the clean calibration maximum against essentially all of them for the tabular heavy-tailed attacker. Appendix D gives the full results.

5.6 The label-aware reference point, and when trimming is unnecessary

Against Label-Trim, which receives the full calibration label set, RCAD is competitive without labels: paired per dataset and seed with 40 trials per cell, it is above Label-Trim for ϵ from 0 to 0.25 on all 3 classical detectors by 0.003 to 0.028 power (t from 2.39 to 4.76 against a critical 2.02 at 39 degrees of freedom) and below by 0.001 to 0.005 at $\epsilon = 0.30$. The advantage is statistically reliable but small in absolute terms, so what it supports is that dropping the labels costs nothing here, not that score-based trimming is the better instrument.

Trimming is also not always needed. Under in-distribution contamination the maximum power difference between RCAD and standard conformal across $\epsilon \in [0, 0.30]$ is 0.000 to 3 decimals at both levels tested, against 0.119 and 0.192 under heavy-tail contamination and 0.123 and 0.206 under adversarial contamination (Table 2): such contamination sits inside the body of the calibration distribution rather than in its upper tail, so its effect on an upper quantile is within sampling noise and there is nothing distinctive at the tail to trim.

6 Discussion

The most useful consequence of this work for a deployed system is that coverage is the wrong thing to monitor on its own: under calibration poisoning it moves in the reassuring direction while the detector stops working. Type-I error at the actual decision threshold, paired with power on a held-out labeled anomaly set however small, separates the two cases that coverage cannot; Section 5.3 shows one diagnostic number reporting 0.9983 for a working detector and a disarmed one at once. The collapse point is also predictable, since Eq. (10) puts it at $\epsilon = \alpha$, so the tolerable contamination fraction can be read off the chosen significance level without running anything.

On the defense, the trimmed quantile is an old estimator and what is new is the analysis of what it does inside a conformal procedure. The correction $\alpha' = \alpha(1 + 2\epsilon)$ cancels the first-order contamination bias exactly, so the drift in Theorem 5 is second order in ϵ, though at our sample sizes the sampling term dominates and the finite-sample content sits in Proposition 8, which also shows a symmetric

trim is the wrong shape for a one-sided decision rule: the configuration exploiting it reaches coverage 0.5002 against a target of 0.900, and the one-sided rule repairing it needs no contamination parameter, which a practitioner cannot look up anyway.

7 Conclusion

Under upper-tail calibration contamination standard split conformal reaches operating level $\min\{(1-\alpha)/(1-\epsilon), 1\}$, so at $\epsilon = \alpha$ it stops raising false alarms while its coverage guarantee still holds, whereas reading the trimmed-quantile threshold as a full-sample order statistic yields a bracket on that level, a power bound, and an expression for attainable coverage that points to a one-sided rule needing no contamination parameter and retaining at least 97% of clean power.

Reproducibility Statement

Proofs of all results in Section 3.3 are given in Appendix A. Appendix E lists the 8 random seeds, the data splits, the detector settings, and the contamination generator. The supplementary material contains the full implementation of both calibration rules, the attack generators, the embedding extraction pipeline, per-trial raw outputs.

Use of Large Language Models

A large language model was used for copy editing of the manuscript, including grammar, word choice, and the wording of captions. It was not used to generate research ideas, to derive or state any mathematical result, to produce or modify any experimental result, or to select references. All theoretical claims were derived by the authors. All references were checked against the primary publication record. The authors take full responsibility for the contents of this submission.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Meshi Bashari, Matteo Sesia, and Yaniv Romano. Robust conformal outlier detection under contaminated reference data. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 3091–3141, 2025.

Stephen Bates, Emmanuel J. Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *Annals of Statistics*, 51(1):149–178, 2023.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning*, pages 1807–1814, 2012.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *ACM SIGMOD International Conference on Management of Data*, pages 93–104, 2000.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022.

Leying Guan and Robert Tibshirani. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society Series B*, 84(2):524–546, 2022.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2nd edition, 2009.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *IEEE Symposium on Security and Privacy*, pages 19–35, 2018.

Rikard Laxhammar and Göran Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74(1): 67–94, 2015.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H. Chen. ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2023.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *IEEE International Conference on Data Mining*, pages 413–422, 2008.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Annals of Probability*, 18(3):1269–1283, 1990.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, volume 2430 of *Lecture Notes in Computer Science*, pages 345–356, 2002.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Marco A. F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pages 4393–4402, 2018.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, volume 30, pages 3517–3529, 2017.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

A PROOFS

Throughout, $A_{(1)} \le \cdots \le A_{(m)}$ are the sorted calibration scores, $\widehat{F}_m$ is their empirical distribution function, F_Q is the distribution function of $A(X)$ under Q in Eq. (2), and F_P is continuous.

A.1 PROOF OF LEMMA 3

By Definition 1 the threshold is $A_{(j)}$ with $j = \lceil (1-\alpha')(m-2k) \rceil + k$, which is an order statistic of the full sample by construction: the trimmed sample consists of positions $k+1$ through $m-k$, and the $\lceil (1-\alpha')(m-2k) \rceil$-th of those is position $\lceil (1-\alpha')(m-2k) \rceil + k$ of the full sample. Substituting $k = \lfloor \epsilon m \rfloor$ and $\alpha' = \alpha(1+2\epsilon)$ and dropping the floor and ceiling,

$$\begin{aligned} \frac{j}{m} &= \frac{(1-\alpha')(m-2k)+k}{m} + O(1/m) = (1-\alpha(1+2\epsilon))(1-2\epsilon) + \epsilon + O(1/m) \\ &= (1-2\epsilon) - \alpha(1+2\epsilon)(1-2\epsilon) + \epsilon + O(1/m) = 1 - \epsilon - \alpha(1-4\epsilon^2) + O(1/m), \end{aligned} \tag{11}$$

using $(1+2\epsilon)(1-2\epsilon) = 1-4\epsilon^2$. The floor contributes at most 1 to k and the ceiling at most 1 to the leading term, so $|j/m - p^\star| \le 2/m$. Differentiating $p^\star$ gives $\partial p^\star / \partial \epsilon = -1 + 8\alpha\epsilon$, which is negative whenever $8\alpha\epsilon < 1$, in particular for all $\epsilon < 1/2$ and $\alpha < 1/4$. □

A.2 PROOF OF LEMMA 4

Since $F_Q = (1-\epsilon)F_P + \epsilon F_{\text{cont}}$, for every t

$$|F_Q(t) - F_P(t)| = \epsilon \, |F_{\text{cont}}(t) - F_P(t)| \le \epsilon, \tag{12}$$

because both distribution functions take values in $[0,1]$. No assumption on P_{cont} is used. The calibration sample is independent and identically distributed from Q, so the Dvoretzky–Kiefer–Wolfowitz inequality in Massart's form (Dvoretzky et al., 1956; Massart, 1990) gives $\sup_t |\widehat{F}_m(t) - F_Q(t)| \le \sqrt{\log(2/\delta)/2m}$ with probability at least $1-\delta$. The triangle inequality gives the stated bound. □

Remark 11. The inequality is applied to the full sample, not to the trimmed one. Trimming selects points by rank, so the retained subsample is neither independent nor independent of the selection rule, and the Dvoretzky–Kiefer–Wolfowitz inequality does not apply to it under any adjustment of constants. Lemma 3 is what makes the detour unnecessary.

A.3 PROOF OF THEOREM 5

Since F_Q is continuous, the calibration scores are almost surely distinct and $\widehat{F}_m(A_{(j)}) = j/m$; where a detector produces ties, randomized tie-breaking restores this. Evaluating Lemma 4 at the data-dependent point $t = \theta = A_{(j)}$, which is legitimate because the bound is uniform in t, gives $|j/m - F_P(\theta)| \le \epsilon + \sqrt{\log(2/\delta)/2m}$ with probability at least $1-\delta$. Combining with $|j/m - p^\star| \le 2/m$ from Lemma 3 and writing $s_{m,\delta} = \sqrt{\log(2/\delta)/2m} + 2/m$,

$$|F_P(\theta) - p^\star| \le \epsilon + s_{m,\delta}. \tag{13}$$

For the upper side, substitute $p^\star = 1 - \epsilon - \alpha(1-4\epsilon^2)$:

$$F_P(\theta) \le p^\star + \epsilon + s_{m,\delta} = 1 - \epsilon - \alpha + 4\alpha\epsilon^2 + \epsilon + s_{m,\delta} = 1 - \alpha + 4\alpha\epsilon^2 + s_{m,\delta}, \tag{14}$$

in which the two ϵ terms cancel identically. For the lower side,

$$F_P(\theta) \ge p^\star - \epsilon - s_{m,\delta} = 1 - 2\epsilon - \alpha(1-4\epsilon^2) - s_{m,\delta} \ge 1 - \alpha - 2\epsilon - s_{m,\delta}, \tag{15}$$

using $\alpha(1-4\epsilon^2) \le \alpha$. □

No inversion of F_P occurs anywhere in this argument. That is deliberate: converting a bound on $|\widehat{F}_m - F_P|$ into a bound on $|\widehat{q} - q|$, where $q = F_P^{-1}(p^\star)$ is the target quantile in score space and $\widehat{q} = \theta$ is its empirical counterpart, requires a positive lower bound on the density near that quantile, which is a different object from an upper bound on the smoothness of the distribution function. An upper bound controls how fast F_P rises and says nothing about how slowly, so if F_P is flat near the target quantile then two empirical distribution functions arbitrarily close in supremum norm can have quantiles at opposite ends of that flat region, with no finite constant relating them. Appendix B reports what this costs on real score distributions.

A.4 Proof of Corollary 7

Let $\theta^* = F_P^{-1}(1-\alpha)$ and $\gamma = 4\alpha\epsilon^2 + s_{m,\delta}$. On the event of Theorem 5, $F_P(\theta) \le 1-\alpha+\gamma$, so $\theta \le F_P^{-1}(1-\alpha+\gamma)$. The survival function $t \mapsto \mathbb{P}[A(X_{\text{anom}}) > t]$ is non-increasing, hence

$$\begin{aligned}\text{Pow}(\theta) &\ge \mathbb{P}\big[A(X_{\text{anom}}) > F_P^{-1}(1-\alpha+\gamma)\big] \\ &= \mathbb{P}[A(X_{\text{anom}}) > \theta^*] - \mathbb{P}\big[A(X_{\text{anom}}) \in \big(\theta^*,\, F_P^{-1}(1-\alpha+\gamma)\big]\big] \\ &= \pi_0 - \mathbb{P}[F_P(A(X_{\text{anom}})) \in (1-\alpha,\, 1-\alpha+\gamma]]\,, \end{aligned} \tag{16}$$

where the last step applies the probability integral transform of F_P, which is valid by continuity. □

A.5 Proof of Proposition 8

Exactly j calibration points lie at or below $\theta = A_{(j)}$. Of these, c are contaminated and $j-c$ are clean, so conditional on the calibration set containing m_0 clean points the fraction of clean points at or below θ is exactly $(j-c)/m_0$. Under Eq. (2) the calibration sample is drawn independently from the mixture, so m_0 is $\text{Binomial}(m, 1-\epsilon)$ and $m_0/m = (1-\epsilon) + O_p(m^{-1/2})$. Substituting and applying Lemma 3,

$$\frac{j-c}{m_0} = \frac{p^\star - c/m}{1-\epsilon} + O(1/m) + O_p(m^{-1/2}). \tag{17}$$

The leading expression is affine in c, and c ranges over $\{0, 1, \ldots, \lfloor \epsilon m \rfloor\}$, giving the stated endpoints. Under a fixed-fraction contamination model, in which exactly $\lfloor \epsilon m \rfloor$ points are replaced, m_0 is deterministic and the $O_p(m^{-1/2})$ term disappears. □

A.6 Proof of Proposition 9

Standard split conformal uses $j_{\text{std}}/m \to 1-\alpha$. Applying Eq. (9) with $p^\star$ replaced by $1-\alpha$ and with $c = 0$, which is what contamination placed entirely above the threshold produces, gives a clean-sample level of $(1-\alpha)/(1-\epsilon)$. A level is a probability, so it is capped at 1, giving Eq. (10). The cap binds when $(1-\alpha)/(1-\epsilon) \ge 1$, that is when $\epsilon \ge \alpha$. At that point θ_{std} lies above the support of the clean score distribution, so the type-I error is 0. Any anomaly distribution supported below θ_{std} then also has power 0; anomalies scoring above the clean support survive, which is the residual power Section 5.1 measures. □

A.7 Breakdown point

Proposition 12 (Breakdown point)**.** *The estimator in Eq.* (3) *has finite-sample breakdown point* $k/m = \lfloor \epsilon m \rfloor / m$ *per tail.*

The threshold in Eq. (3) is a function of order statistics with index between $k+1$ and $m-k$. Corrupting at most k points in a single tail changes only which points occupy the trimmed positions, leaving the retained middle block unchanged in value. With $k+1$ corruptions in one tail the corrupted mass reaches the retained block and the estimator can be driven to an arbitrary value. The breakdown point is therefore k/m. □

B Conditioning of the inversion step

Write $\kappa = L_F / f_{\min}$ for the ratio that an inversion from level space to score space introduces, with L_F a Lipschitz constant for F_P and $f_{\min}$ a density floor near the target quantile. Estimated on held-out clean scores at the $\alpha = 0.10$ operating point across 4 datasets, median of 3 seeds, κ runs from 4.74 to 9.79 for Isolation Forest, 3.81 to 25.55 for one-class SVM, and 8.28 to 53.87 for Local Outlier Factor. On mammography it is unbounded, and literally so: that benchmark carries 30.5 % exactly duplicated normal rows, so the Local Outlier Factor reachability density diverges at zero distance, its scores run a 99th percentile of 2.13 against a maximum of 2.4×10^9, and the measured $f_{\min}$ is below 10^{-100} on every seed. A score-space bound carrying a coefficient $(1+\kappa)\epsilon$ returns 1.08 at the median κ with $\epsilon = 0.10$, a deviation larger than the range of a probability, so the resulting coverage guarantee

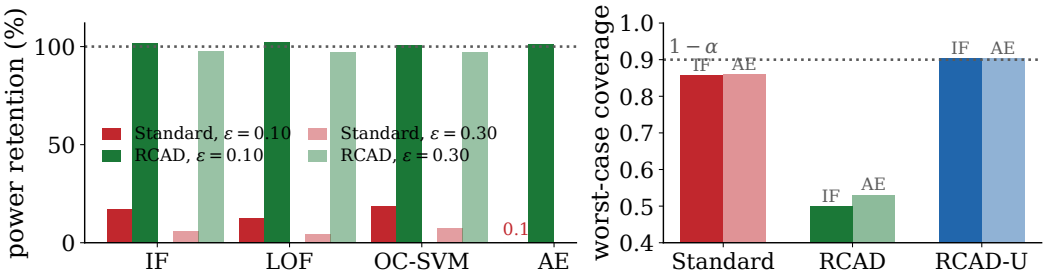

Figure 2: Left: power retention, meaning power at contamination level ϵ as a percentage of the same configuration's clean power, for 4 detector families under adversarial calibration poisoning; the dotted line marks full retention, and standard conformal retains 0.1% on the autoencoder (AE) at $\epsilon = 0.10$. Right: worst coverage reached anywhere in 2,400 configurations spanning 3 detectors, 5 datasets, 8 seeds, 4 contamination placements and ϵ from 0 to 0.30, against the target $1 - \alpha = 0.90$ dotted. Solid bars are Isolation Forest, faded bars the autoencoder.

falls below zero and certifies nothing; on mammography it is undefined. This is why Theorem 5 is stated in level space.

C Additional Experimental Results

Predicted against measured clean-data type-I error. Lemma 3 predicts the type-I error on clean data as $1 - j/m$ from the trimming parameters alone. Table 1 compares the prediction against measurement on Isolation Forest at $m = 200$ with a true contamination rate of 0.

Table 1: Type-I error on clean calibration data, predicted from Lemma 3 against measured, Isolation Forest, $m = 200$, $\alpha = 0.10$, 8 seeds.

Assumed ϵ	0.00	0.10	0.20	0.30
Predicted $1 - j/m$	0.1000	0.1950	0.2800	0.3600
Measured	0.0975	0.1976	0.2881	0.3659
Deviation	0.0025	0.0026	0.0081	0.0059

Power retention and worst-case coverage. Figure 2 gives the detector sweep and the worst-case coverage summary discussed in Sections 5.2 and 5.4.

Misspecification of the contamination budget. Figure 3 gives the full grid summarized in Section 5.4.

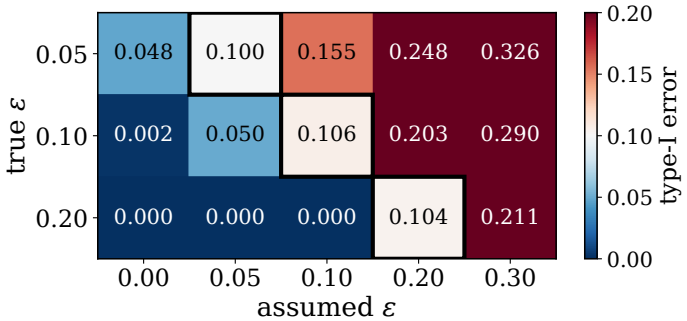

Figure 3: Type-I error of RCAD as the assumed contamination budget is swept independently of the true one, Isolation Forest, $m = 200$, $\alpha = 0.10$, 8 seeds. Boxed cells are the correctly specified diagonal. Underspecification moves left and is conservative for validity; overspecification moves right and inflates the error rate.

Contamination regimes.

Table 2: Maximum power difference, RCAD minus standard conformal, over $\epsilon \in [0, 0.30]$ on the synthetic benchmark, by contamination regime and level, 8 seeds. The defense is specific to contamination placed in the upper tail.

Contamination regime	$\alpha = 0.05$	$\alpha = 0.10$
In-distribution	0.000	0.000
Heavy-tail	0.119	0.192
Adversarial	0.123	0.206

Coverage under clean calibration. On clean calibration data with $m = 600$, empirical coverage deviates from $1 - \alpha$ by at most 0.026 over $\alpha \in \{0.01, 0.05, 0.10, 0.15, 0.20\}$, and with an assumed $\epsilon = 0$ the rule of Definition 1 reduces to standard split conformal, so RCAD and standard conformal agree exactly there.

The second-order cap. Table 5 tests the $4\alpha\epsilon^2$ term of Theorem 5 against measured coverage growth over the $\epsilon = 0$ baseline, under in-distribution contamination on the synthetic benchmark at $m = 200$.

Table 3 gives coverage and power for the 6 embedding configurations of Figure 4, with clean power reported so that any retention ratio can be recomputed. We do not quote a retention ratio for the 20 Newsgroups Mahalanobis cell, whose clean power of 0.159 is too small for a ratio to be stable.

Table 3: Coverage and power at a calibration contamination rate of 0.20 on frozen embeddings, $\alpha = 0.10$. Each cell reports coverage and power separated by a slash. The clean-power column is measured at $\epsilon = 0$, and c1, c9 and c7 name the class held out as the anomaly class.

Representation	Score	Clean	Standard	RCAD	RCAD-U
CLIP CIFAR-10 c1	kNN	0.915	0.989 / 0.513	0.863 / 0.946	0.961 / 0.801
CLIP CIFAR-10 c1	Mahalanobis	0.763	0.972 / 0.453	0.845 / 0.858	0.935 / 0.634
CLIP CIFAR-10 c9	kNN	0.935	0.989 / 0.495	0.863 / 0.949	0.958 / 0.798
CLIP CIFAR-10 c9	Mahalanobis	0.732	0.974 / 0.433	0.854 / 0.841	0.942 / 0.623
MiniLM 20 News c7	kNN	0.373	0.937 / 0.259	0.792 / 0.612	0.921 / 0.301
MiniLM 20 News c7	Mahalanobis	0.159	0.920 / 0.139	0.748 / 0.370	0.917 / 0.143

Reference-set size. Table 4 gives the ablation summarized in Section 5.4.

Table 4: Sensitivity of RCAD-U to the size of the clean reference sample. Isolation Forest, upper-tail contamination at a true rate of approximately 0.05, $\alpha = 0.10$, 8 seeds. Power rises monotonically with reference size; the tail-mass estimate is noisy at small n_{ref} and is not itself monotone. Standard conformal never reads the reference and is flat by construction.

Reference size n_{ref}	40	80	160	320
$\widehat{\epsilon}_{\mathrm{up}}$ (true ≈ 0.05)	0.043	0.031	0.047	0.052
RCAD-U power	0.267	0.290	0.325	0.526
Standard conformal power	0.065	0.065	0.065	0.065

A recalibration pipeline. To check whether the tail-mass estimator of Eq. (4) survives a setting where the reference is not guaranteed clean, we simulate periodic recalibration on a temporally split stream, drawing the reference either from the same window as the calibration set or from the preceding one. When the reference is drawn from a window carrying its own contamination, the estimator recovers the *difference* between calibration and reference contamination rather than the absolute rate: on mammography with a true injected rate of 0.10, same-window and prior-window references return 0.011 and 0.013. This is the expected behavior of a two-sample excess statistic and it marks the condition RCAD-U needs, namely a reference that is cleaner than the calibration set, not merely another sample from the stream.

D REPRESENTATION-SPACE RESULTS

Repeating the experiment on frozen CLIP image embeddings of CIFAR-10 and frozen MiniLM sentence embeddings of 20 Newsgroups, at a calibration contamination rate of 0.20, the ordering found on tabular data holds in all 6 cells: standard conformal at the highest coverage and lowest power, RCAD at the lowest coverage and highest power, and RCAD-U between them. Across the 4 CIFAR-10 configurations standard conformal retains 53 to 59% of its clean power against 102 to 115% for RCAD. Figure 4 and Table 3 give the numbers, with clean power reported so that any retention ratio can be recomputed; we do not quote a ratio for the 20 Newsgroups Mahalanobis cell, whose clean power of 0.159 is too small for one to be stable.

Two differences from the tabular setting are informative about the attack rather than the defense. First, standard conformal retains far more power here, 0.139 to 0.513 against near zero on tabular data, and Eq. (9) says why: complete shutdown requires contamination above the clean support, and in embedding space the attack does not get there. The fraction of injected points scoring above the clean calibration maximum is 0.141 and 0.115 on CIFAR-10 class 1, 0.244 and 0.060 on class 9, and 0.005 and 0.001 on 20 Newsgroups, with the median contaminated point at the 99th and 96th percentile of clean calibration scores on CIFAR-10 and the 85th and 63rd on 20 Newsgroups, whereas the tabular heavy-tailed attacker places essentially all of its mass beyond that support. Second, text is harder than vision for both rules: RCAD's coverage deficit roughly triples, from 0.037 and 0.055 on CIFAR-10 to 0.108 and 0.152 on 20 Newsgroups, and on 20 Newsgroups with Mahalanobis scores RCAD-U is indistinguishable from standard conformal, which the placement measurements above account for. Results for CIFAR-10 classes 1 and 9 agree to 3 decimals on coverage and within 0.02 on power, so the vision result is stable across the choice of anomaly class.

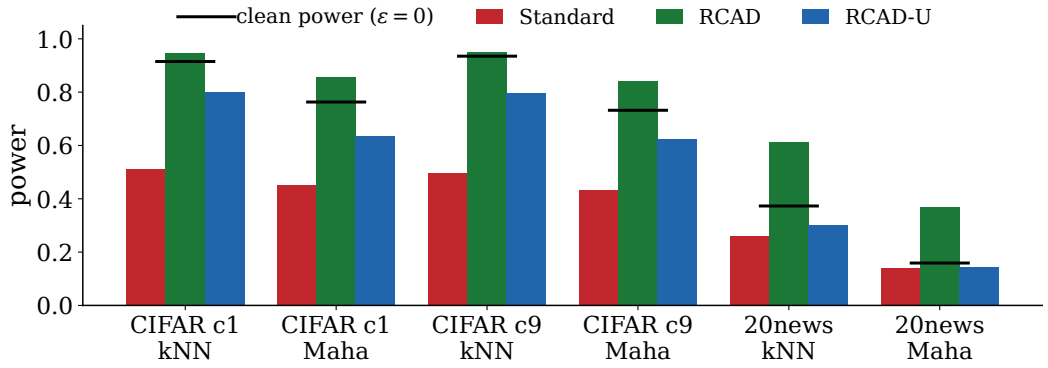

Figure 4: Power at a calibration contamination rate of 0.20 on frozen CLIP embeddings of CIFAR-10 (classes 1 and 9 as the anomaly class) and frozen MiniLM embeddings of 20 Newsgroups (class 7), with k nearest neighbor and Mahalanobis scores, $\alpha = 0.10$. Black dashes mark each configuration's clean power at $\epsilon = 0$.

Table 5: Measured coverage growth over the $\epsilon = 0$ baseline against the predicted $4\alpha\epsilon^2$, synthetic benchmark, $m = 200$, $\alpha = 0.10$, 8 seeds. The measured growth tracks $4\alpha\epsilon^2$ to within 0.0012 up to $\epsilon = 0.20$, slightly exceeding it at $\epsilon = 0.10$ and 0.20, which Theorem 5 permits because its bound also carries $s_{m,\delta}$; at $\epsilon = 0.30$ the measurement falls well below.

ϵ	0.05	0.10	0.20	0.30
Measured	$+0.0007$	$+0.0043$	$+0.0172$	$+0.0199$
Predicted	0.0010	0.0040	0.0160	0.0360
Residual	-0.0003	$+0.0003$	$+0.0012$	-0.0161

E EXPERIMENTAL DETAILS

Seeds and splits. All results are means over the 8 seeds $\{42, 1, 7, 13, 21, 100, 200, 300\}$, with seed-level standard deviations reported in the raw outputs. For each seed the data is split into train and test, and the training normals are split into a fitting set and a calibration set using a calibration fraction $\gamma = 0.30$. Performance varies by under 5% across $\gamma \in [0.2, 0.4]$, and is stable across $\alpha \in [0.05, 0.15]$.

Contamination generator. The white-box candidate pool contains resampled real anomalies together with Gaussian draws at 8σ from the calibration mean along random unit directions. For each seed and contamination level we draw $\max(20k, 200)$ candidate points, where $k = \lfloor \epsilon m \rfloor$. Heavy-tailed candidates are Gaussian draws at 8σ from the calibration mean along a random unit direction, together with up to $5k$ resampled real anomalies. Candidates are scored under the detector appropriate to the attacker arm and the top k by nonconformity are submitted. Placement variants restrict the submitted points to a target quantile band of the clean score distribution, which is how the 4 placements in Section 5.4 are generated. The attack uses score values only and assumes no gradient access.

Score conventions. All detectors are converted to the convention that larger values are more anomalous, which requires negating the decision function of `IsolationForest` and `OneClassSVM`, and the sample scores of the Local Outlier Factor. Applying this inconsistently inverts the reported rankings of the affected baselines.

Cost. Training time scales as $O(n \log n)$, from 0.11 seconds at $n = 500$ to 0.17 seconds at $n = 10{,}000$ on commodity hardware, dominated by the detector fit. The full experimental suite completes in under 40 minutes on a single machine.