COUNTERFACTUAL FRAGILITY CERTIFICATES FOR TABULAR CLASSIFICATIONS: GUARANTEES AND LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

A high-confidence tabular prediction need not survive unreliable input fields. We study protocol-relative fragility: loss of support for the original prediction under a declared finite collection of field failures. Redundant fields can mask one-step effects, while partial degradation can differ from deletion. Counterfactual Fragility Certificates (CFC) combine a sequential removal path, recomputed after every step, with graded tests of every group. An original-class margin records collapse even after the predicted class changes. A reconstructable record exposes the tests, decisions, and witnesses behind a fixed fragility dominance score (FDS). We distinguish record verification from worst-case robustness, causal recourse, and prediction correctness. A conditional greedy bound and counterexamples delineate generic search limits; a model-specific affine solver certifies minimum removal depth. We evaluate a released implementation on seven OpenML datasets, eight classifier families, and three seeds, with paired row-level outputs and reference checks. On the in-family endpoint, CFC-FDS achieves 0.7743 AUROC versus 0.7741 for prespecified, query-matched random search with the same three-term score; the paired difference is +0.0001 (95% interval [-0.0054, 0.0063]). On withheld donor/quantization mechanisms, the difference is -0.0040 ([-0.0089, 0.0004]). These comparisons do not establish a consistent advantage for adaptive removal over query-matched random search across both endpoints. A separate five-dataset restoration study measures 2.36 net errors averted per 100 cases at 10% review using a validation-trained utility ranker, versus 2.61 for raw confidence.

1 INTRODUCTION

A confident tabular prediction can depend on unreliable fields. Two measurements may mask each other's importance until both fail; another field may become misleading when slightly degraded. This motivates two questions: *what declared field failures can overturn this prediction, and would checking the record improve the decision?* Confidence on the intact input does not answer the first; fragility alone does not answer the second.

We address these questions with a fixed classifier and field-level stress tests. All encoded columns of a raw field form one group. A protocol specifies replacements, degradation levels, and a test budget before evaluation. "Counterfactual" means evaluating modified inputs; it does not assert a causal intervention or actionable recourse. The target of the audit is loss of support for the *original prediction*. Whether that prediction is wrong is a separate question requiring labels.

Why this design? CFC combines two complementary branches. A sequential removal branch recomputes each remaining group's effect after every removal, because redundancy can change which field matters next. A graded branch tests *every* group at several attenuation, dropout, and noise levels, because partial degradation can reveal failures missed by complete replacement. Both branches track the original class: a confident switch to another class must count as lost support. The resulting certificate stores the prescribed tests, outputs, removal decisions, and failure witnesses. A verifier reconstructs the protocol to detect missing or misreported evidence. Figure 1 shows this workflow.

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.

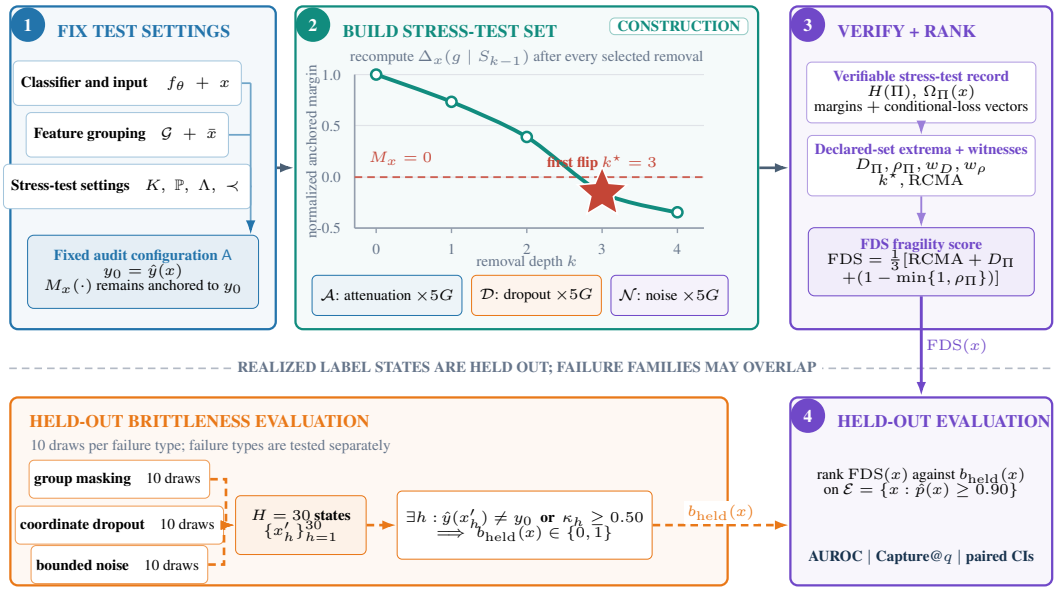

Figure 1: **Why and how CFC tests high-confidence predictions.** The original-class margin is recomputed after each removal; the curve is a schematic, not a measured result. All groups receive graded tests. The record verifies the declared computation, while FDS summarizes it for ranking. The orange branch shows the primary in-family endpoint: ten draws per mechanism, with a flip or at least 50% margin collapse marking a positive. Separate label states prevent realized-outcome reuse; shared failure families still limit transfer claims. Withheld mechanisms and observed errors are evaluated separately.

What distinguishes CFC? CFC combines search, graded testing, and a reconstructable audit record. It builds on existing ideas: sufficient input subsets already use backward selection (Carter et al., 2019), and removal benchmarks evaluate feature importance (Samek et al., 2017; Hooker et al., 2019). CFC keeps the classifier fixed and verifies its finite audit, including the comparisons behind each removal; it does not certify invariance under untested failures. For known affine scores, a separate solver certifies minimum removal depth over all baseline-replacement subsets; arbitrary greedy search has no such guarantee. Our counterexamples show why.

From an audit to a review decision. The fixed fragility dominance score (FDS) summarizes how early support collapses, its largest collapse, and the smallest tested cost of a flip. These describe sensitivity, not the value of correcting a record. We therefore also study *CFC-U*, a separate ranker trained on validation examples to predict the net error reduction from restoring a record. This separates three questions: is the audit faithful, does its score rank fragility, and does the review action help?

Our released experiments compare fragility rankings on seven datasets under matched query budgets and measure restoration corrections *and harms* on five new datasets. They do not establish consistent superiority over matched comparators; raw confidence remains competitive. The supported contributions are an inspectable finite audit, scoped exact removal guarantees, and reproducible evidence separating verification, ranking, and intervention benefit.

2 Related Work

Uncertainty and selective prediction. Calibration evaluates probability reliability (Guo et al., 2017; Widmann et al., 2019); uncertainty estimates can degrade under distribution shift (Ovadia et al., 2019). Failure prediction uses confidence or representation-based signals to rank likely errors (Hendrycks & Gimpel, 2017; Jiang et al., 2018; Corbière et al., 2019), while selective classification evaluates prediction risk at reduced coverage (Geifman & El-Yaniv, 2017; 2019; Traub et al., 2024).

CFC instead ranks sensitivity to a specified perturbation protocol. Its primary AUROC and capture metrics use brittleness labels rather than classification errors.

Attribution, removal, and counterfactual explanations. Attribution assigns importance to observed inputs (Lundberg & Lee, 2017), and counterfactual explanation searches for prediction-changing inputs (Karimi et al., 2020; Pawelczyk et al., 2022). MoRF and ROAR evaluate explanations by removal, with ROAR retraining models after deletion (Samek et al., 2017; Hooker et al., 2019). Sufficient input subsets use backward selection (Carter et al., 2019); Anchors finds local sufficient rules with probabilistic guarantees (Ribeiro et al., 2018). CFC instead verifies a declared finite audit, combining an original-class margin, graded all-group tests, and a reconstructable record. Perturbation choices affect interpretation (Yeh et al., 2019; Janzing et al., 2020); our groups and replacements are testing conventions rather than causal explanations.

Robustness certificates. Adversarial certificates prove prediction invariance within a norm ball; randomized smoothing certifies an ℓ_2 radius (Cohen et al., 2019). CFC certifies something different and narrower: that a recorded set of responses to a prescribed, finite collection of field failures is complete and correct. It does not imply that the prediction is correct or stable under untested failures.

Tabular robustness. Tabular robustness depends on preprocessing, and tree ensembles remain strong (Arik & Pfister, 2021; Gorishniy et al., 2021; Grinsztajn et al., 2022; Hollmann et al., 2025). Shift and adversarial benchmarks test population-level or threat-model robustness (Gardner et al., 2023; Simonetto et al., 2024). CFC instead audits individual predictions under declared field-level failures.

3 METHOD

Overview. Given a fixed classifier, input, and protocol, CFC (1) constructs a removal path and graded tests, (2) records and verifies their responses, and (3) summarizes fragility with FDS. CFC-U learns a separate review score. Known affine models additionally permit exact removal search.

3.1 SETTING AND FAILURE ASSUMPTIONS

Let $x \in \mathbb{R}^d$ be a preprocessed feature vector with class label $y \in \{1, \ldots, C\}$. A trained classifier f_θ outputs probabilities $p_\theta(c \mid x)$; the prediction is $\hat{y}(x) = \arg\max_c p_\theta(c \mid x)$ and its confidence is $\hat{p}(x) = \max_c p_\theta(c \mid x)$. CFC holds θ fixed; its audit measures susceptibility to field failure. The original benchmark focuses on high-confidence predictions; the restoration study also evaluates all predictions.

Feature groups. A feature group $g \subseteq \{1, \ldots, d\}$ contains coordinates that fail together; the partition $\mathcal{G} = \{g_1, \ldots, g_G\}$ has G groups. In our benchmarks each group holds every transformed column derived from one raw field, including all one-hot columns of a categorical field. *Why:* a field that goes missing loses all its encoded columns at once, so whole-field removal should replace its encoding together. Graded tests can still produce invalid categorical encodings; they are stress probes. Grouping is fixed and carries no causal guarantee.

Protocol. The locked protocol is $\Pi = (\phi, \mathcal{G}, \bar{x}, K, \mathbb{P}, \mathbf{\Lambda}, \prec, \eta)$: preprocessing ϕ, groups, a training-derived replacement vector $\bar{x}$, maximum removal depth $K \leq G$, failure transformations $\mathbb{P}$ with severity grids $\mathbf{\Lambda}$, a tie-breaking order $\prec$, and numerical tolerance $\eta = 10^{-12}$. An audit is the configuration $\mathsf{A} = (x, f_\theta, \Pi)$, identified by the protocol hash $H(\Pi)$. *Why:* locking the protocol before test evaluation makes its choices inspectable. The hash identifies metadata, not trustworthy provenance or selection time; model and preprocessor files have separate checksums. Supplement S1 gives full definitions.

3.2 MEASURING THE FRAGILITY OF ONE PREDICTION

Fix the original predicted class $y_0 = \hat{y}(x)$. A perturbed input $\tilde{x}$ is a modified copy of the same sample. With clipped, renormalized probabilities $\tilde{p}$ and centered log-probabilities $\tilde{z}_c(\tilde{x}) =$

$\log \tilde{p}_c(\tilde{x}) - C^{-1} \sum_{a=1}^{C} \log \tilde{p}_a(\tilde{x})$ (Supplementary Section S1), the *anchored margin* is

$$M_x(\tilde{x}) = \tilde{z}_{y_0}(\tilde{x}) - \max_{c \neq y_0} \tilde{z}_c(\tilde{x}). \tag{1}$$

It is positive when y_0 uniquely wins, zero at a tie, and negative when a competitor has the larger score. Flips at ties use the fixed class order. *Why anchored:* if the reference class were re-chosen after each perturbation, a prediction that had already flipped would show a large positive margin for its new class and appear robust. *Why log-probabilities:* all eight model families expose probabilities, so the log probability ratio provides a common definition across model families; it does not guarantee equal calibration. A flip, $\hat{y}(\tilde{x}) \neq y_0$, is not necessarily an error; that requires the true label.

Sequential group removal. Let $\mathcal{R}(x, S)$ replace the coordinates of groups S by $\bar{x}$. The conditional support loss of group g after removing S is

$$\Delta_x(g \mid S) = M_x(\mathcal{R}(x, S)) - M_x(\mathcal{R}(x, S \cup \{g\})). \tag{2}$$

Starting from $S_0 = \varnothing$, for $k = 1, \ldots, K$ CFC selects

$$g_k = \arg \max_{g \in \mathcal{G} \setminus S_{k-1}} \Delta_x(g \mid S_{k-1}), \qquad S_k = S_{k-1} \cup \{g_k\}, \qquad x^{(k)} = \mathcal{R}(x, S_k), \tag{3}$$

choosing the first group under $\prec$ whose loss is within η of the maximum. *Why recompute:* $\Delta_x(g \mid S)$ depends on S. A field whose evidence is duplicated elsewhere has small loss while its partner is present and large loss once the partner is removed. Recomputation can expose that change once a partner is selected, but cannot guarantee that a hidden interaction is reached. The path flip depth $k^\star$ is the first k with $\hat{y}(x^{(k)}) \neq y_0$, or $K + 1$ if none; it need not equal the globally smallest failing subset, which Section 5 measures.

Graded degradation. In parallel with the removal path, each group is tested on the original input: values are moved toward $\bar{x}$ (attenuation), coordinates are partially replaced (dropout), or numerical values receive bounded noise, each at five severities. *Why:* real fields often degrade rather than disappear, and a model can be far more sensitive to a noisy value than to its replacement by a typical one. *Why every group:* the greedy path explores only one sequence; testing every group covers each declared operator–severity pair. Other directions, severities, or mechanisms remain untested. The declared test collection is

$$\Omega_\Pi(x) = \{x^{(0)}, \ldots, x^{(K)}\} \cup \{\mathcal{P}_{g,\lambda}(x) : g \in \mathcal{G}, \mathcal{P} \in \mathbb{P}, \lambda \in \Lambda_{\mathcal{P}}\}, \tag{4}$$

whose $15G$ graded tests keep distinct identities and costs even when their inputs coincide. Coverage means every prescribed test is present, not every possible failure. Auxiliary candidate, coordinate, and direction probes are retained for verification; the summaries below range over $\Omega_\Pi(x)$, not every auxiliary model query.

Summaries and review score. For a tested input v, the normalized margin collapse is $r_x(v) = \min\{1, [(M_x(x) - M_x(v))/\max\{M_x(x), 10^{-8}\}]_+\}$, set to 1 after a flip. Normalizing by the original margin makes collapse comparable across predictions of different confidence. Along the path, $\bar{r}_k = \max_{0 \leq j \leq k} r_x(x^{(j)})$ prevents a later recovery from hiding an earlier collapse. We define

$$\mathrm{RCMA}(x) = \frac{1}{K} \sum_{k=1}^{K} \bar{r}_k, \qquad D_\Pi(x) = \max_{v \in \Omega_\Pi(x)} r_x(v). \tag{5}$$

RCMA is large when support is lost *early* on the removal path; D_Π is the worst collapse over $\Omega_\Pi(x)$. With protocol costs $c_\Pi(x^{(k)}) = k/K$ and $c_\Pi(\mathcal{P}_{g,\lambda_\ell}(x)) = \ell/5$ for severity index ℓ, the minimum flip cost is

$$\rho_\Pi(x) = \min_{v \in \Omega_\Pi(x),\ \hat{y}(v) \neq y_0} c_\Pi(v), \qquad \rho_\Pi(x) = 6/5 \text{ if no test flips.} \tag{6}$$

Costs are positions in the protocol, not failure probabilities. Changing K, the grouping, or a severity grid changes their meaning; FDS values should be compared under the same locked protocol. The three summaries capture distinct aspects of fragility: how quickly support erodes (RCMA), how far it falls (D_Π), and how little failure is needed to reverse the prediction (ρ_Π). The review score averages them with fixed, equal weights:

$$\mathrm{FDS}(x) = \tfrac{1}{3}\big[\mathrm{RCMA}(x) + D_\Pi(x) + 1 - \min\{1, \rho_\Pi(x)\}\big] \in [0, 1]. \tag{7}$$

Why fixed weights: FDS provides a fully specified summary without fitting to evaluation labels; equal weighting is a convention, not an optimality claim. CFC-U below retains these components but learns a separate score for a different target: restoration benefit.

3.3 WHAT THE CERTIFICATE GUARANTEES

The certificate stores the protocol hash, every test with its classifier response, the full conditional-loss vector at every removal step, the summaries, and *witnesses*: the tests attaining maximum collapse (w_D) and minimum flip cost (w_ρ), or an explicit no-flip result. A verifier re-runs the protocol to obtain a reference record $R_\Pi(x; f_\theta)$ and compares it field by field, rather than trusting the submitted summaries.

Why this matters: replaying only the tests a producer chose to report cannot reveal an omitted test. Reconstructing the *required* tests can. Verification requires access to the fixed input, classifier, and preprocessing and repeats the construction cost; it is not a succinct proof or a check of data authenticity. Under exact comparison, the verifier accepts a certificate if and only if it equals the reference record (Proposition S1). Acceptance therefore rules out omitted tests, incorrect removal choices, and misstated summaries. With tolerance η, discrete fields must match exactly and numerical fields within η. This contract concerns execution; prediction correctness and ranking quality require separate evidence.

Finally, for compatible test collections, union yields $D_{1\sqcup 2} = \max(D_1, D_2)$ and $\rho_{1\sqcup 2} = \min(\rho_1, \rho_2)$ (Proposition S2). *Why this matters:* audits can be extended with new failure types without discarding or weakening previously recorded failures. Proofs and algorithms are in Supplementary Sections S1–S2.

3.4 STRONGER CERTIFICATES UNDER AFFINE STRUCTURE

For known affine class scores $z_c(x) = w_c^\top x + b_c$, a fixed baseline permits exact grouped-removal optimization. For each rival $c \neq y_0$, let $m_c = z_{y_0}(x) - z_c(x)$ and $a_{cg} = \sum_{j \in g}(w_{y_0 j} - w_{cj})(x_j - \bar{x}_j)$. Write $a^+_{c(1)} \geq \cdots \geq a^+_{c(G)}$ for sorted positive parts.

Proposition 1 (Exact affine removal envelope)**.** *In real arithmetic, for every budget k,*

$$\min_{|S| \leq k} \left[z_{y_0}(\mathcal{R}(x, S)) - \max_{c \neq y_0} z_c(\mathcal{R}(x, S)) \right] = \min_{c \neq y_0} \left[m_c - \sum_{j=1}^{k} a^+_{c(j)} \right]. \tag{8}$$

The first budget at which a rival's residual becomes negative (or zero with tie priority over y_0) is the global minimum flip depth for this fixed-baseline removal family. Its sorted positive groups supply a witness.

Maximizing additive loss for each rival and commuting the minima gives this result in $O(Cd + CG \log G)$ operations. It applies established tractable contrastive explanation ideas (Cooper & Marques-Silva, 2023; El Harzli et al., 2023). It covers every removal subset of the supported affine model; it does not cover graded noise, causal changes, or an arbitrary nonlinear classifier. The formula uses affine scores, not clipped log probabilities. The implementation has an independent rational-arithmetic checker for minimum depth in the mathematical model represented by the stored coefficients and inputs. Deployed floating-point execution is distinct. Supplement S6 gives the proof, a multiclass greedy counterexample, and the generic black-box boundary.

3.5 RANKING FOR A SPECIFIED REVIEW ACTION

A fragile prediction is not necessarily worth reviewing: replacing a corrupted record can correct an error, leave it unchanged, or introduce one. For observed and restored records, define the signed benefit

$$g = \mathbb{1}\{f(x_{\text{observed}}) \neq y\} - \mathbb{1}\{f(x_{\text{clean}}) \neq y\} \in \{-1, 0, 1\}.$$

Here $+1$ is a correction, 0 no change in correctness, and -1 a harm. *CFC-U* is a regressor fitted only on validation examples to predict this benefit from confidence deficit, entropy, negative margin, and the three FDS components. At test time it receives the observed record's scores; a fixed review budget selects its highest predictions. Restoration then reveals the clean record and the original classifier runs again. If the conditional expectations were known, this selection would maximize expected net corrections by the standard exchange argument; the fitted regressor does not inherit that guarantee. The same learner defines Confidence-U from the three confidence features and MC-U from those features plus the three matched MC components. Test selection never sees the clean

record, fault identity, true label, or restoration benefit. This extension requires labeled restoration examples; it does not turn FDS itself into a calibrated probability of error.

4 Experimental Protocol

Frozen design and audited population. We use Adult, Bank, Credit-G, Default, Electricity, HELOC, and Covertype from fixed OpenML versions. For each seed in $\{17, 29, 43\}$, stratified splits allocate 70%/15%/15% to training, validation, and test. Training is capped at 100,000 uniformly sampled rows. Eight fixed model families are logistic regression, random forest, extra trees, histogram gradient boosting, MLP, XGBoost, LightGBM, and CatBoost. Hyperparameters are fixed; the MLP's early stopping uses training data only. Numeric imputation/scaling and categorical missing-token one-hot encoding are fitted on training rows. Each group comprises a raw field's transformed columns. CFC uses their training mean as replacement, training scales and ranges, and $K = \min(10, G)$. Supplement S3 lists full configurations.

We sample up to 500 test rows with original confidence at least 0.90, uniformly before labels or fragility scores are computed. A unit is a dataset–model–seed combination. Units with fewer than 50 sampled eligible rows are excluded. Each endpoint additionally requires at least five positives and five negatives; exclusions apply identically to all methods. The run audits 65,419 sample–model–seed pairs across 156 retained units. The primary endpoint includes 152 units across 7 datasets. Endpoint-specific denominators and every exclusion are released; absence of an estimable endpoint is not a zero score.

Pilot and provenance. The protocol and code were committed before the seed-101 Adult/Credit-G pilot. Its checks concerned prevalence, agreement, and runtime, not method ranking. All thresholds were retained for the final run; the pilot is excluded from reported final estimates. Using different seeds does not prevent individual rows from overlapping across pilot and final splits. The release stores split IDs, frozen models, fitted preprocessors, protocols, sampled certificate JSONs, row-level scores, labels, environment, and checksums. Three certificates per retained unit are reconstructed against the slower reference; all 468 checks agree. The first full attempt stopped on numerical reconstruction failures; a documented adapter normalization and deterministic forest-prediction repair preceded a complete rerun. No scientific thresholds changed.

Four endpoint definitions. For every eligible row, the original class and margin remain fixed. Labels are generated before scores with a separate random seed namespace. The primary *in-family* endpoint uses 30 draws: ten group masks, ten coordinate-dropout states, and ten numeric-noise states. A row is brittle if any draw flips its prediction or reduces the original-class margin by at least 50%. Independent realizations avoid outcome reuse, but the construction and label mechanisms remain related.

The *withheld-mechanism* endpoint uses ten donor-row group substitutions and ten training-decile midpoint quantizations, each affecting one to three groups. Neither mechanism is in the construction protocol; their mathematical support need not be disjoint from its states. *Missing-pattern stress* applies ten training-observed NaN masks to the raw row and then invokes the fitted preprocessor, using the same flip/collapse rule. This is a generated failure test using empirical patterns, not prospective incident validation; datasets without observed NaN masks have no such endpoint. Special source codes are not silently reinterpreted as NaNs. *Induced error* restricts to originally correct eligible predictions and marks any misclassification among the in-family or withheld-mechanism states. It assumes the true label is invariant to corruption; donor substitution can violate that assumption. None of these endpoints alone establishes benefit from human review.

Exploratory observed-error check. After inspecting stress results on six completed datasets, we separately froze an analysis of errors in the unperturbed original predictions. It reuses all retained audit cohorts and scores, with the same minimum class counts and aggregation. Unlike induced error, its labels require no corruption-invariance assumption. It is exploratory and does not replace the primary endpoint.

Primary comparator and controlled ablations. Prespecified MC-FDS receives the same original-state, coordinate-probe, noise-direction, and all-group graded tests as CFC. It spends

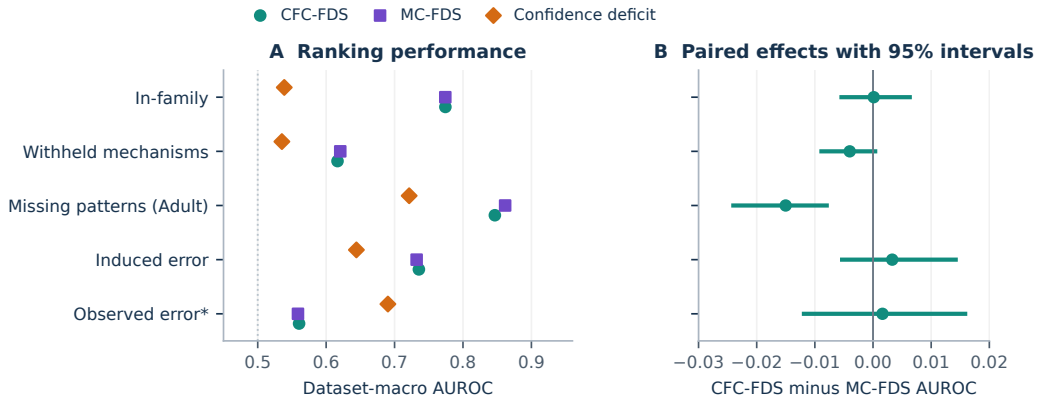

Figure 2: **Fragility ranking and error ranking answer different questions.** Left: dataset-macro AUROC for CFC-FDS, MC-FDS, and confidence deficit. Right: paired CFC-minus-MC effects with the frozen 95% bootstrap intervals. Positive effects favor CFC. Missing-pattern stress covers Adult alone; observed error is exploratory. Endpoint cohorts differ.

$B = \sum_{k=1}^{K}(G - k + 1)$ evaluations on random removal subsets: cardinality uniformly sampled from $1, \ldots, K$, then a uniform subset, independently with replacement. Both methods use exactly

$$N_{\text{eval}} = 1 + B + d + 2d_{\text{num}} + 15G$$

single-row model evaluations (batched in practice). These counts exclude label generation and verification. For MC-FDS, the removal-area term averages, over budgets $k = 1, \ldots, K$, the greatest sampled collapse among subsets of size at most k, using zero when none was sampled. It uses the same three-term head as CFC. This compares different search geometries, not algorithms differing only in one isolated operation. MC-D versus CFC-D separately compares maximum-collapse summaries.

The fixed-order ablation ranks initial single-field losses once, then follows that order. It evaluates the same candidate budget at subsequent steps but does not use those losses to change the order. The graded field and FDS formula are unchanged: this is an information-use ablation, not an optimized low-cost fixed-order implementation. Path-only FDS discards the graded summaries while retaining the same queried outputs. Other baselines are confidence deficit, entropy, negative margin, maximum one-step loss, 20 random-order RCMA trajectories, and a validation-derived permutation order. Their costs differ and are not called query-matched.

Metrics and inference. AUROC is computed within each evaluable unit, averaged equally over units within a dataset, and then equally over represented datasets. Capture@q is brittle recall among the top $\lceil qn \rceil$ rows for $q \in \{.05, .10, .20\}$, with original-ID score ties. Its per-unit ceiling is $\min\{1, \lceil qn \rceil / B_u\}$ for B_u brittle rows; the exact random-review reference is $\lceil qn \rceil / n$. Macro capture cannot be checked using pooled prevalence alone. Paired 95% intervals use 10,000 dataset-first, then within-dataset unit bootstrap replicates, retaining identical paired rows and labels. This resamples a small set of datasets with fixed within-unit rows and does not capture all training/population uncertainty. The primary contrast is CFC-FDS minus MC-FDS; secondary contrasts are exploratory without multiplicity correction.

5 Results

Primary result. In-family AUROC is 0.7743 for CFC-FDS and 0.7741 for MC-FDS (paired difference +0.0001, 95% interval [-0.0054, 0.0063]); this interval does not resolve a difference. Figure 2 compares the principal methods and their paired effects; the supplement reports every method and per-dataset result. Every result is generated directly from the released row outputs.

Transfer and outcome-linked stress. Withheld-mechanism AUROC is 0.6166 for CFC-FDS and 0.6206 for MC-FDS (paired difference -0.0040, 95% interval [-0.0089, 0.0004]); this interval does

Table 1: Exploratory component contrasts on the identical in-family units. Positive values favor the first method.

Paired contrast	AUROC change	95% interval
CFC-FDS − Fixed-order FDS	+0.0073	[0.0019, 0.0131]
CFC-FDS − Path-only FDS	+0.0114	[-0.0017, 0.0345]
CFC-D − MC-D	-0.0368	[-0.0774, -0.0022]
CFC-FDS − MC-FDS	+0.0001	[-0.0054, 0.0063]
CFC-FDS − CFC-RCMA	-0.0015	[-0.0191, 0.0245]

not resolve a difference. Missing-pattern stress AUROC is 0.8466 for CFC-FDS and 0.8616 for MC-FDS (paired difference -0.0150, 95% interval [-0.0240, -0.0080]); this interval favors MC-FDS. Induced-error AUROC is 0.7356 for CFC-FDS and 0.7323 for MC-FDS (paired difference +0.0033, 95% interval [-0.0053, 0.0142]); this interval does not resolve a difference. The sets of evaluable units differ by endpoint. Accordingly, endpoint means are not a paired cross-endpoint comparison, and sparse coverage limits transfer claims even when a within-endpoint difference is positive. Confidence is a substantial comparator rather than a disposable sanity check; all its endpoint results are retained in the tables.

Actual prediction errors. On the exploratory unperturbed-error endpoint (131 units), CFC-FDS AUROC is 0.5605, MC-FDS 0.5588, and confidence deficit 0.6903. The CFC-minus-confidence difference is -0.1298 (95% interval [-0.1889, -0.0722]). This is a separate question from predicting sensitivity to field failure; its analysis was defined during execution and is not confirmatory evidence.

Which components contribute? Table 1 compares the fixed-order, path-only, and maximum-only alternatives on the same primary endpoint units. RCMA alone attains 0.7758 AUROC; the full-head difference is -0.0015, with paired interval [-0.0191, 0.0245]. Thus these data do not establish that the three-term head is superior to its path-area component. These effects apply to the declared protocol and model collection. They do not establish that a greedy search is best over all possible failure families, groupings, or query allocations.

Search quality. Exhaustive subset search was feasible for 960 sampled cases from Adult and Electricity; greedy and exact flip-depth values agreed in 97.1%. Among 549 cases with a flip within budget, greedy required greater depth or found no flip in 28. These exact-feasible cases are a restricted subset. Supplement S2 gives conditional greedy bounds and counterexamples; their assumptions are not established for the fitted nonlinear classifiers.

Exact affine certificates and sensitivity. The affine solver covers 7,825 retained logistic-regression audits across 19 units. It agrees with all 120 applicable exhaustive-search checks and all 3,259 returned witnesses flip the saved model. Greedy uses greater depth in 18 Covertype cases, including 6 missed within the ten-group budget. All 7,825 minimum-depth and no-flip records pass this exact-rational reconstruction. The post-primary threshold/draw sensitivity grid and exact-arithmetic validation are fully reported in Supplement S6. They strengthen the search characterization without establishing a ranking advantage.

Review prioritization and cost. Primary-endpoint capture at 20% review is 36.3% (1.81× the exact-random reference); it measures brittle-case retrieval, not prevented errors. Median unit-amortized construction time is 1.05 ms, with at most 1424 model evaluations per audit. These single-run Apple M4 timings include concurrent contention; complete query and timing definitions are in Supplement S3.

5.1 A SEPARATE INTERVENTION EXPERIMENT ON FIVE NEW DATASETS

We freeze a new study on Spambase, Phoneme, Steel Plates Fault, Wilt, and Banknote Authentication (OpenML IDs 44, 1489, 1504, 40983, 1462), outside the original benchmark. Three fixed classifiers (LR, RF, HGB) and three new seeds use 60/20/20 train/validation/test splits. Validation

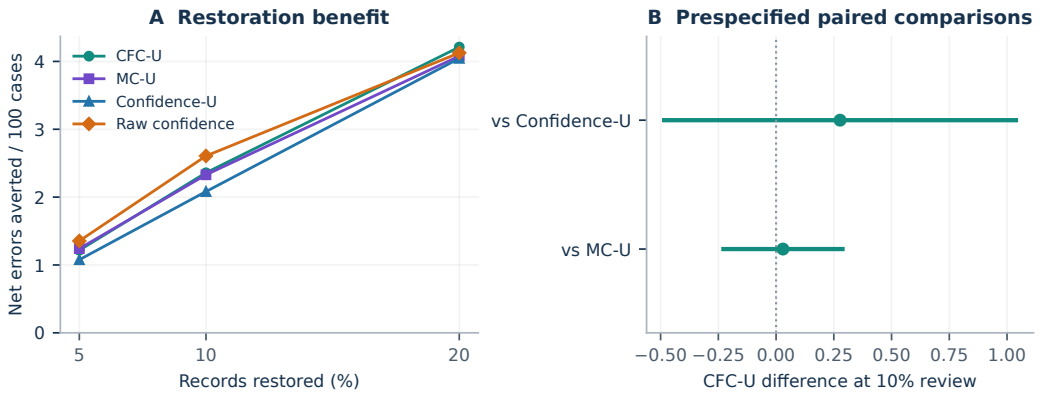

Figure 3: **Measure what review changes, including harm.** Left: net errors averted under three fixed record-restoration budgets on five new datasets. Right: the two primary CFC-U contrasts at 10%, with multiplicity-adjusted intervals. U denotes validation-trained utility ranking; raw confidence remains an important, cheaper baseline. These results concern the controlled restoration resource.

faults are mean replacement and Gaussian noise; test faults use donor substitution and stuck-at-training-quantile values, with 20% unchanged records in both channels. The original case label is fixed by the measurement-fault simulation. Each selected review restores the entire clean record at unit cost; the classifier then predicts again. This is an idealized restoration resource, not a prospective deployment study.

The shared regression architecture is fixed before test evaluation. Primary evaluation retains all predictions; high-confidence results are secondary. The primary budget is 10%, with equal averaging over units within datasets, then datasets. For the two prespecified CFC-U comparisons, paired dataset-level t intervals use a Bonferroni adjustment (97.5% per interval). All outcomes, including harmful and no-op reviews, remain in the ledger; Supplement S7 gives the complete design and results. Across 45 configurations and 16,776 test instances, CFC-U averts 2.359 net errors per 100 cases at 10% review, compared with 2.082 for Confidence-U, 2.329 for MC-U, 2.607 for raw confidence, and 0.981 for random review. The CFC-U differences are +0.277 versus Confidence-U ([-0.485, +1.039]) and +0.030 versus MC-U ([-0.228, +0.288]); these multiplicity-adjusted intervals do not establish superiority. Raw confidence also remains a stronger point-estimate baseline.

5.2 Limitations and Responsible Use

A correct certificate can encode an unrealistic failure model. Mean replacement, fractional categorical states, and source pre-encoded groups need not describe valid joint observations. FDS values and costs are comparable only under the same protocol. Sparse endpoint coverage and chosen model/data populations limit transfer; the bootstrap does not capture all shared-row and training dependence. The affine guarantee is for a specified mathematical model, while the restoration study assumes access to a clean record. Neither establishes causal relevance or prospective deployment benefit. Use certificates to inform review, not automatic adverse decisions; all limitations and exclusions remain in the supplement.

6 Conclusion

CFC distinguishes faithful recording, exact restricted search, and useful review. Reconstruction and the affine solver address the first two. Across the original ranking and new restoration studies, neither fixed FDS nor CFC-U establishes consistent superiority over matched baselines. The contribution is an inspectable, executable account of certificate guarantees and decision-level limits, including net benefit and restoration harm. Competitive ranking and operational validation remain open.

Reproducibility Statement

The supplement gives the protocol, exact algorithms, complete proof, configurations, pilot decisions, exclusions, and generated detailed tables. The accompanying local release contains frozen source and protocol commits, dependency versions, row-level endpoint CSVs, models, preprocessors, split IDs, sampled verified certificates, and result-generation scripts. All reported empirical numbers are generated from these outputs. Reference and benchmark tests pass; sampled reconstruction checks cover 468 audits in the original experiment, plus 45 independent checks in the restoration study. Exact-rational replay checks all 7,825 affine records.

Ethics Statement

Several benchmarks (Adult, Bank, Credit-G, Default, HELOC) concern consequential decisions about individuals. CFC does not change predictions; it is intended to direct human attention to fragile ones. Its score must not be used as an automatic basis for adverse action, and the fields it flags reflect model dependence rather than causal or normative relevance.

AI Use Statement

We used Codex to assist with manuscript formatting. The authors take responsibility for the final content, including all mathematical claims and experimental results.

References

Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6679–6687, 2021. doi: 10.1609/aaai.v35i8.16826.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? understanding black-box decisions with sufficient input subsets. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 567–576. PMLR, 2019.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019.

Martin C. Cooper and João Marques-Silva. Tractability of explaining classifier decisions. *Artificial Intelligence*, 316:103841, 2023. doi: 10.1016/j.artint.2022.103841.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2902–2913, 2019.

Ouns El Harzli, Bernardo Cuenca Grau, and Ian Horrocks. Cardinality-minimal explanations for monotonic neural networks. In *Proceedings of IJCAI*, pp. 3677–3685, 2023. URL https://www.ijcai.org/proceedings/2023/0409.pdf.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53385–53432, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pp. 507–520, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9737–9748, 2019.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2907–2916. PMLR, 2020.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 895–905. PMLR, 2020.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *Proceedings of the Twenty Fifth International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4574–4594. PMLR, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11491. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

Thibault Simonetto, Salah Ghamizi, and Maxime Cordy. Tabularbench: Benchmarking adversarial robustness for tabular deep learning in real-world use cases. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2323–2347, 2024.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12236–12246, 2019.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10965–10976, 2019.