Set-Level Risk Enrichment Does Not Establish Reliable Pointwise Rankings: Wrong-Label Causal Vulnerability in Two Tabular Foundation-Model Readouts

Anonymous authors
Paper under double-blind review

Abstract

Mechanistic studies of tabular foundation models can identify context examples that strongly influence prediction and use these signals to construct targeted poisoning attacks. Attack effectiveness, however, does not answer two stronger questions: whether high-influence positions are more vulnerable to wrong labels than matched-random positions under the same corruption budget, and whether the same importance score can rank this damage pointwise. We test both questions across two mechanistically characterized readouts, using deletion of the same rows to separate the harm of retaining wrong supervision from removing the samples themselves. Retaining wrong labels is more harmful than deletion for the high-influence sets selected by both readouts. For TabPFNv2's attention-vote readout, the combined 20-dataset analysis shows a 2.62 accuracy-point excess $[0.78, 4.65]$ for high-influence over class-matched random corruption; the same excess is Inconclusive when re-aggregated on the shared 9-dataset $\times$ 3-split scope. In a separate 9-dataset pointwise analysis, the same pre-corruption importance shows weak correspondence with single-point damage: macro Spearman correlation is -0.069 for NLL damage and $+0.011$ for accuracy damage. TabICLv2's prototype readout establishes neither stable set-level risk enrichment nor reliable pointwise ranking. Across the tested scopes, set-level enrichment does not by itself establish reliable pointwise risk ranking; the two properties require separate validation.

1 Introduction

Tabular foundation models predict new rows from a labelled context in a single forward pass (Müller et al., 2022; Hollmann et al., 2023; 2025). Two tabular foundation models have been characterized mechanistically. TabPFNv2 emits an attention-weighted vote over context labels at a late layer, and TabICL classifies by comparing a query against class-mean prototypes (Qu et al., 2026; Biloš et al., 2026). These readouts are part of the released models rather than interchangeable heads (Biloš et al., 2026). Prior work therefore lets us say how a context point participates in a prediction, in each family and with a concrete score.

Mechanism-targeted attacks show that these readouts can be exploited: corrupting mechanism-selected positions degrades both models substantially (Biloš et al., 2026). Attack effectiveness leaves three questions unresolved: whether high-influence positions are more vulnerable than matched-random positions under the same corruption budget, whether the harm comes from retaining wrong supervision rather than removing the same row, and whether the underlying importance score ranks the damage of individual label errors. Answering them requires counterfactual comparisons that hold the corruption budget and the label-transition rule fixed.

Three questions operationalize the correspondence between the tested mechanism-derived prediction importance and wrong-label causal vulnerability, from the coarsest to the finest. The wrong-supervision question: does keeping a wrong label cost more than deleting the same row? The set-enrichment question: is corrupting the high-influence set more harmful than corrupting a class-stratified matched-random set of the same size? The pointwise question: does a pre-corruption,

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.

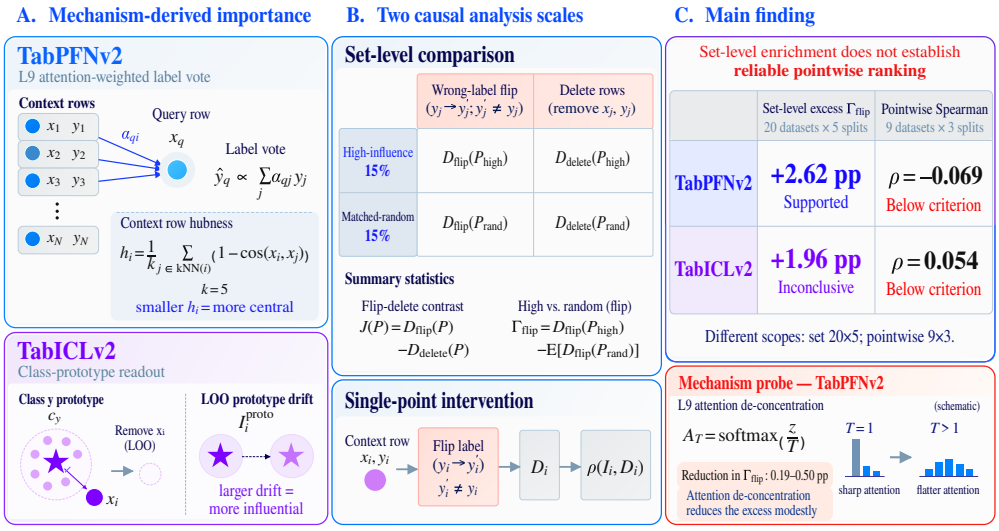


Figure 1: Counterfactual design and main finding. (A) Mechanism-derived importance for the two readout families: TabPFNv2's L9 attention-weighted label vote with context-row hubness h_i, and TabICL's class-prototype readout with the leave-one-out prototype drift I_i^{proto}. (B) The two causal analysis scales: at the set level wrong-label flips are contrasted against deletions for high-influence and matched-random positions, summarized by the flip–delete contrast $J(P)$ and the high-vs-random excess Γ_{flip}; at the single-point level each context point's pre-corruption importance is correlated with its realized damage D_i. (C) Main finding across the two analyses: the attention-vote readout shows a supported set-level excess (+2.62pp, 14/20) while neither readout's query-label-free importance reaches the reliable-ranker criterion pointwise on the 9-dataset × 3-split study ($\rho = -0.069$ for the oriented hub importance $H_i = -h_i$ and $\rho = +0.054$ for the prototype drift), and the L9 attention de-concentration probe ($A_T = \text{softmax}(z/T)$) reduces the excess by only 0.19–0.50pp, and the set-level and pointwise analyses cover different experimental scopes.

mechanism-aligned importance score rank the damage caused by flipping each context point individually? These questions probe distinct properties of each model's importance score.

The two analyses give different evidence patterns. Keeping a wrong label is more harmful than deleting the same row, and at the set level the attention-vote readout's influence measure identifies a higher-risk set: corrupting high-influence positions costs 2.62pp [0.78, 4.65] more than matched-random corruption across the 20 datasets, although the effect differs between the original and expansion cohorts (Inconclusive on the original 8 datasets, Support on the 12-dataset expansion). The same readout's own importance score does not meet the reliable-ranker criterion for pointwise ranking ($\rho = -0.069$, $2/9$ datasets direction-consistent), and the pointwise conclusion persists when damage is measured in accuracy ($+0.011$) rather than NLL. The prototype readout establishes neither a stable set-level excess nor a reliable pointwise ranking.

We separate set-level risk enrichment from pointwise rankability using matched-random position contrasts at a fixed corruption budget, a flip–delete contrast that quantifies the additional harm of retaining wrong supervision relative to deleting the same row, and exhaustive single-point interventions over every context position. Figure 1 summarizes the design and the distinction between the two analyses.

2 Related Work

Mechanism-targeted poisoning. Biloš et al. (2026) characterize the readout of each released model and use that characterization to construct attacks: hub poisoning for the attention-vote family and rank deformation for the prototype family, both of which degrade predictions substantially. That work establishes that mechanism-targeted poisoning can cause harm. What we add is a distinction

between two properties of the same importance measure: whether it enriches a vulnerable set, and whether it ranks the damage of individual label errors. Prior work did not isolate the excess of high-influence positions over a matched-random set under a fixed budget, or the harm of keeping a wrong label relative to deleting the same row. A parallel robustness line perturbs the features of tabular foundation models at test time, asks whether these models can generate transferable evasion, and studies in-context defenses (Djilani et al., 2025); the corruption we study leaves features and rows untouched and changes only the supervision attached to context labels.

Label noise and in-context corruption. Learning under noisy labels and adversarial corruption has a long history, including robust training, label-flip attacks, backdoors and data poisoning (Biggio et al., 2012; Han et al., 2018; Goldblum et al., 2023; Song et al., 2022). In-context learning raises the same question in a different setting: replacing demonstration labels with random ones leaves accuracy surprisingly close to the ground-truth-label condition in the language-model settings studied (Min et al., 2022), while a closer look at input-label correspondence finds that whether ground-truth labels matter varies with the experimental configuration (Yoo et al., 2022). Both lines vary what the demonstration labels contain, without a position-matched counterfactual for the wrong labels themselves. For TabPFN specifically, Hu & Ghelichi (2026) study robustness to random, position-unselective label noise and report that internal mechanisms adapt under it. That work varies the corruption rate; our primary estimand fixes the budget and varies only the *location* of the wrong supervision, which makes matched-random positions rather than a noise-rate parameter the relevant counterfactual.

Data influence and causal importance. The interpretability literature has long cautioned that quantities associated with prediction formation need not coincide with causal importance under intervention, an argument made most sharply for attention (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Data-influence and attribution methods (Koh & Liang, 2017; Ghorbani & Zou, 2019; Pruthi et al., 2020) answer a different question as well: how useful a training point is, rather than how much damage its corrupted supervision causes. How much a point contributes to a clean prediction, how useful a training point is in general, and how much damage its wrong label causes are therefore three distinct estimands, and a mechanism-aligned set can enrich the third without the same score ranking it pointwise.

3 Counterfactual Study Design and Causal Quantities

3.1 Provenance and study scope

We use a common counterfactual design for both readouts, and we document the study scope and provenance below. The original TabPFNv2 study followed a recorded analysis plan; Appendix A documents the supplied record and its provenance limitations. Separate analysis plans were used for the TabICLv2 extension, the twelve-dataset expansion, and the intervention probe.

3.2 Models and readouts

TabPFNv2 is the official pretrained classifier (Hollmann et al., 2025), used without weight modification under its fp32-deterministic pipeline (8 ensemble members; details in Appendix A.2). Its readout is the deep-block (L9) row-attention-weighted label vote (Biloš et al., 2026): a surrogate vote built on that mechanism reproduces the model's accuracy within a few pp. TabICL is the official `tabicl` package with its v2 checkpoint (Qu et al., 2026); its readout is the class-prototype rule over final representations (Biloš et al., 2026). We include a TabICL release in downstream causal analyses when the prototype readout reaches an agreement of at least 0.80 with the native decoder in each of its 15 query-prediction checks; the included and excluded releases are reported with the single-point results (§5).

3.3 Corruption construction and position sets

The set-level cohort is 20 public classification datasets: the 8 original datasets plus 12 more drawn from the public OpenML-CC18 pool under a pre-fixed eligibility rule. Each dataset is split stratified 80/20 into context/query under 5 split seeds (context ≤ 800, query ≤ 200, features ≤ 48; when a cap binds, the truncation keeps the class composition). The pointwise study is 9 datasets $\times$ 3 splits (the

original 5 plus 4 selected from the expansion cohort under the recorded selection rule), shared by both model arms and reported as the pointwise main analysis. The set-level study covers 20 datasets $\times$ 5 splits, so the two analyses cover different scopes (§8). The corruption budget is $\lfloor 0.15 \times n \rfloor$ context points, fixed across all conditions and all cohorts.

High-influence sets are defined per mechanism on the *clean* context. For TabPFNv2, the hub set P_{hub} takes the $\lfloor 0.15 \times n \rfloor$ rows with smallest L9 cosine centrality $h_i = \text{mean}_{j \in k\text{NN}(i), k=5}(1 - \cos(x_i, x_j))$ over L9 representations (ties by ascending sample ID; the hub phenomenon in high-dimensional nearest neighbors (Radovanović et al., 2010)); intuitively, smaller h_i identifies rows embedded more centrally within dense local neighborhoods of the L9 representation space. For pointwise ranking analyses we orient this quantity as $H_i = -h_i$, so that larger values consistently denote greater mechanism-derived importance and the pointwise score uses the same direction as the set-level selector above. For TabICLv2, the high-influence set takes the top-15% of the prototype-drift score I^{TabICL} (ties by sample ID). The selectors are mechanism-specific by construction, and each model is compared only with its own matched-random control. Matched-random sets $P^{(r)}_{\text{rand}}$ (r=1, …, 10) satisfy $|P^{(r)}_{\text{rand}}| = |P_{\text{high}}|$ and $P^{(r)}_{\text{rand}} \cap P_{\text{high}} = \varnothing$, drawn class-stratified to match the high set's original-class counts (when a class has too few remaining rows, the draw takes the largest feasible count and is flagged). Flip rule: each selected label is flipped uniformly to one of the other $K - 1$ classes; the high-influence set and every matched-random set use the same uniform-other-class rule with deterministic per-set random streams. Deletion baseline: the same positions are removed from the context. Write $\mathcal{C}$ for the clean context, $\mathcal{C}'_{\mathcal{P}}$ for the flipped context and $\mathcal{C} \setminus \mathcal{P}$ for the deleted context of any position set $\mathcal{P}$.

3.4 Decomposition quantities, decision thresholds and the three-state criterion

With primary metric fixed to Accuracy, define for any position set $\mathcal{P}$:

$$D_{\text{flip}}(\mathcal{P}) = \text{Perf}(\mathcal{C}) - \text{Perf}(\mathcal{C}'_{\mathcal{P}}), \tag{1}$$

$$D_{\text{delete}}(\mathcal{P}) = \text{Perf}(\mathcal{C}) - \text{Perf}(\mathcal{C} \setminus \mathcal{P}), \tag{2}$$

$$J(\mathcal{P}) = \text{Perf}(\mathcal{C} \setminus \mathcal{P}) - \text{Perf}(\mathcal{C}'_{\mathcal{P}}) = D_{\text{flip}}(\mathcal{P}) - D_{\text{delete}}(\mathcal{P}), \tag{3}$$

$$\Gamma_{\text{flip}} = D_{\text{flip}}(P_{\text{high}}) - \tfrac{1}{10} \textstyle\sum_r D_{\text{flip}}(P^{(r)}_{\text{rand}}), \tag{4}$$

$$\Gamma_{\text{delete}} = D_{\text{delete}}(P_{\text{high}}) - \tfrac{1}{10} \textstyle\sum_r D_{\text{delete}}(P^{(r)}_{\text{rand}}), \tag{5}$$

$$\Gamma_J = J(P_{\text{high}}) - \tfrac{1}{10} \textstyle\sum_r J(P^{(r)}_{\text{rand}}). \tag{6}$$

$J(\mathcal{P})$ is the extra net performance harm of keeping wrong labels relative to deleting those samples outright; no "pure label effect" interpretation is made. The statistical unit is the dataset: within a dataset we average over splits and random sets, then take the unweighted macro mean over datasets; confidence intervals are dataset-level paired bootstrap 95% CIs (seed 20260908, 10,000 resamples (Efron & Tibshirani, 1993)). The practical-effect threshold is fixed at 1pp (the prespecified practical difference of the original design (Jaeschke et al., 1989)) with ROPE= $\pm$0.5pp (Kruschke, 2018; Lakens et al., 2018). The three-state criterion (Support / Reject / Inconclusive, mutually exclusive): Support requires point estimate at or above the practical-effect threshold, a CI lower bound above zero and at least two thirds of the datasets direction-consistent; Reject is a 95% CI inside the ROPE or a CI upper bound < 0; otherwise Inconclusive. A composite claim is SUPPORT only if every positive sub-claim is. Inconclusive denotes insufficient evidence under this rule rather than equivalence. For the single-point correspondence, the reliable-ranker criterion is macro Spearman $\rho \geq 0.20$ with a confidence interval whose lower bound exceeds zero and at least two thirds of the datasets direction-consistent. Failure to meet the reliable-ranker criterion does not imply zero association.

3.5 Pre-corruption signals

For the single-point diagnostic we record, per context point i and before any corruption, three TabPFNv2 signals that are query-label-free and pre-corruption: hubness H_i (defined above), max clean query attention $M_i = \max_q \bar{A}^{\text{clean}}_{qi}$, and mean clean query exposure $\bar{E}_i = \frac{1}{Q} \sum_q \bar{A}^{\text{clean}}_{qi}$ ($\bar{A}$ averages attention over heads and ensemble members). These scores may use the context labels, which

Table 1: Main cross-model table. (a) Set-level causal risk by cohort: J_{hi} (keeping wrong labels vs. deleting those rows) and Γ_{flip} (high-influence vs. matched-random excess) in accuracy points, with dataset-level paired bootstrap 95% CI and direction counts in parentheses; verdicts use the three-state criterion, and the original-cohort values reproduce the original analysis. (b) Pointwise correspondence on the 9-dataset × 3-split study: macro Spearman ρ (direction counts in parentheses) against the primary D_{NLL} and against D_{Acc}, which shares the accuracy scale of (a), with per-cohort CIs in Table 18. Panels (a) and (b) summarize different experimental scopes.

(a) Set-level causal risk by cohort

Model	J_{hi} (pp)	Γ_{flip} (pp)	J verdict	Γ verdict
Original cohort (8)				
TabPFNv2	$+3.43$ [$+2.06$, $+5.00$] (8/8)	$+0.49$ [-0.33, $+1.23$] (5/8)	Support	Inconclusive
TabICLv2	$+4.08$ [$+2.30$, $+6.19$] (8/8)	$+1.86$ [-1.46, $+4.99$] (6/8)	Support	Inconclusive
Expansion cohort (12)				
TabPFNv2	$+5.94$ [$+3.00$, $+9.00$] (10/12)	$+4.03$ [$+1.19$, $+7.12$] (9/12)	Support	Support
TabICLv2	$+4.28$ [$+2.44$, $+6.37$] (12/12)	$+2.02$ [-1.82, $+5.31$] (9/12)	Support	Inconclusive
Combined cohort (20)				
TabPFNv2	$+4.93$ [$+3.04$, $+6.90$] (18/20)	$+2.62$ [$+0.78$, $+4.65$] (14/20)	Support	Support
TabICLv2	$+4.20$ [$+2.87$, $+5.74$] (20/20)	$+1.96$ [-0.66, $+4.34$] (15/20)	Support	Inconclusive

(b) Pointwise correspondence, 9×3

Signal (damage)	Original 5	Expansion 4	Combined 9
TabPFNv2 H (D_{NLL})	-0.009 (2/5)	-0.144 (0/4)	-0.069 (2/9)
TabPFNv2 H (D_{Acc})	-0.004 (3/5)	$+0.030$ (2/4)	$+0.011$ (5/9)
TabICLv2 I^{proto} (D_{NLL})	$+0.036$ (4/5)	$+0.077$ (2/4)	$+0.054$ (6/9)
TabICLv2 I^{proto} (D_{Acc})	$+0.022$ (3/5)	-0.007 (2/4)	$+0.009$ (5/9)

are ordinary model inputs, but never the query ground-truth labels. For TabICLv2 the query-label-free pre-corruption signal is the prototype-drift score I_i^{TabICL}: the class-prototype readout itself is established by Biloš et al. (2026), while the leave-one-out drift is our operationalization of it, not an importance method released with TabICL. A query-label-aware offline oracle diagnostic, the first-order readout-gradient score I^{grad}, is defined in Appendix C, analyzed separately (§5), and is used only as an offline reference because it requires query ground-truth labels.

For each context row i we form the single-flip counterfactual $\mathcal{C}_i'$ that flips only the label of i (uniformly to one of the other $K-1$ classes) and re-run inference on the unchanged query set. The single-point damages are

$$D_i^{\text{Acc}} = \text{Acc}(\mathcal{C}) - \text{Acc}(\mathcal{C}_i'), \qquad D_i^{\text{NLL}} = \text{NLL}(\mathcal{C}_i') - \text{NLL}(\mathcal{C}),$$
$$D_i^{\text{prob}} = \tfrac{1}{Q}\sum_q \big(p_{\mathcal{C}}(y_q) - p_{\mathcal{C}_i'}(y_q)\big), \tag{7}$$

each evaluated on the query set, so that larger values mean more damage. NLL is computed on the model's probability output with the standard machine-epsilon clipping of the log; per dataset, the points of all splits are pooled into one Spearman correlation per signal and damage, and cells whose signal or damage is undefined (constant or non-finite) are excluded from the macro mean.

4 Causal Vulnerability at the Set Scale

We answer the first two questions on the mechanism-defined high-influence set of each readout, under the fixed 15% corruption budget, for both models simultaneously. Table 1 summarizes all headline results; Table 3 reports every quantity separately for the original, expansion-only and combined cohorts. The statistical unit is the dataset throughout, with splits aggregated inside a dataset and dataset-level paired bootstrap intervals.

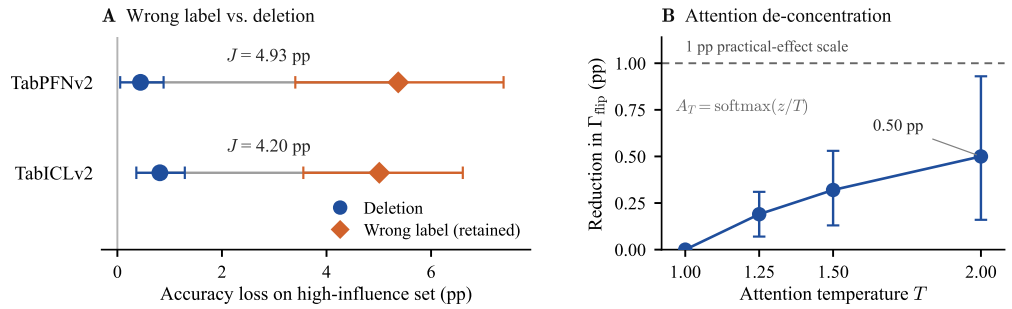


Figure 2: Wrong-supervision harm and a mechanism-matched probe. (A) Accuracy loss from retaining wrong labels versus deleting the same mechanism-defined high-influence rows, aggregated over the combined 20-dataset cohort. Points show macro means and error bars show dataset-level bootstrap 95% confidence intervals. Retaining wrong labels is substantially more harmful than deleting the same rows for both readouts, with flip–delete contrasts of 4.93 pp for TabPFNv2 and 4.20 pp for TabICLv2. (B) Reduction in TabPFNv2's set-level excess under attention de-concentration on the combined 20 datasets $\times$ 5 splits, implemented as $A_T = \mathrm{softmax}(z/T)$. Error bars show 95% confidence intervals for the reduction relative to $T{=}1$. The effect is directionally consistent but remains below the 1 pp practical-effect scale.

4.1 Wrong supervision relative to deletion

Across every cohort, retaining wrong supervision is more harmful on average than deleting the same rows. On the combined twenty-dataset cohort the flip–delete decomposition is $J = 4.93$pp $[3.04, 6.90]$ (18/20) for TabPFNv2 and $J = 4.20$pp $[2.87, 5.74]$ (20/20) for TabICLv2 (Figure 2A). Both meet the Support criterion, with the same qualitative pattern in the original and expansion cohorts (smaller and larger point estimates respectively; Table 3). J measures the additional harm of retaining corrupted supervision relative to deleting the same rows; it is not a pure label effect, and the same qualitative flip–delete pattern is present at the matched-random positions as well.

4.2 High-influence versus matched-random positions

TabPFNv2. Across all 20 datasets the high-vs-matched-random excess for the attention-vote readout is $\Gamma_{\text{flip}} = 2.62$pp $[0.78, 4.65]$ (14/20), Support. This Support result is specific to the prespecified Accuracy outcome; secondary metrics do not support a metric-independent enrichment claim (Appendix G). That excess is heterogeneous across provenance-defined cohorts: the original 8 remain Inconclusive at 0.49pp $[-0.33, 1.23]$ (5/8), whereas the expansion 12 reach Support at 4.03pp $[1.19, 7.12]$ (9/12). The deletion-side contrasts are near zero on all three cohorts (0.01pp, 0.08pp, 0.05pp), with confidence intervals spanning zero. The set-level result shows that the selector enriches higher-risk positions; §5 asks whether the underlying continuous score also ranks individual damage.

TabICLv2. For the prototype readout the combined excess is $\Gamma_{\text{flip}} = 1.96$pp $[-0.66, 4.34]$ (15/20), Inconclusive, and the same verdict holds on the original cohort (1.86pp, 6/8) and the expansion cohort (2.02pp, 9/12); the interval spans zero on every cohort, so no stable set-level excess is established.

Heterogeneity. $14/20$ datasets agree in direction, satisfying the two-thirds criterion, and the per-dataset effect is heterogeneous: Γ_{flip} spans -3.00pp to $+13.49$pp. Exploratory dataset-level correlates were weak in every case we measured ($|\text{Spearman}| \leq 0.38$ for TabPFNv2 and at most 0.53 over both models; Appendix B). These exploratory correlates do not explain the observed heterogeneity, whose source remains unresolved.

5 Pointwise Correspondence

We next test whether the continuous importance scores rank single-point damage on the shared 9-dataset × 3-split pointwise study, using the aggregation and the reliable-ranker criterion defined in §3.

5.1 TabPFNv2: broad-cohort enrichment and weak pointwise ranking

On the combined 9-dataset cohort the oriented hub importance $H = -h$ reaches $\rho_{\text{NLL}} = -0.069$ $[-0.126, -0.016]$ with $2/9$ datasets direction-consistent, a macro mean far below the 0.20 threshold: the reliable-ranker criterion is not met. The modest negative macro association is concentrated in the expansion subset. On the original five datasets $\rho = -0.009$ $[-0.052, +0.034]$ $(2/5)$, while on the four expansion datasets it is a modest negative association that still falls short of the threshold, -0.144 $[-0.192, -0.088]$ $(0/4)$. The pointwise conclusion persists when damage is measured in accuracy ($\rho_{\text{Acc}} = +0.011$ $[-0.029, +0.055]$, $5/9$). A single-point flip leaves the query-set accuracy unchanged for most flips (Table 30), so NLL is the primary pointwise damage measure and accuracy the scale-matched secondary check. The auxiliary attention signals do not recover the ranking on the primary damage measure (M: $+0.015$; $\bar{E}$: -0.009).

5.2 TabICLv2: neither property is established

The prototype readout establishes neither a stable set-level excess (§4) nor a reliable pointwise ranking: leave-one-out prototype drift gives $\rho_{\text{NLL}} = +0.054$ $[-0.013, +0.136]$ $(6/9)$ on the primary damage measure (auxiliary D_{prob}: $+0.012$) and $\rho_{\text{Acc}} = +0.009$ $[-0.037, +0.046]$ $(5/9)$ on the accuracy scale.

Among the official TabICL releases, v1 and v2 satisfy the readout-agreement criterion and both miss the pointwise ranking criterion, whereas v1.1 falls below the agreement threshold; details are in Table 9.

5.3 Query-label-aware offline oracle

Among the signals evaluated on the original five datasets, the first-order diagnostic I^{grad} (Appendix C) is the only one that meets the reliable-ranker criterion on the primary D_{NLL} outcome; it uses query ground-truth labels, so it is an offline reference rather than a deployable score. There it reaches $\rho_{\text{NLL}} = 0.387$ $[0.254, 0.541]$ $(5/5)$, $\rho_{\text{prob}} = 0.231$ $[0.152, 0.293]$ $(5/5)$, and $\rho_{\text{Acc}} = 0.226$ $[0.167, 0.290]$ $(5/5)$: all three damage measures pass the reliable-ranker criterion. The oracle shows substantially stronger rank correspondence, but this comparison changes both the score construction and access to query labels and therefore does not isolate the contribution of query-label information alone.

5.4 Synthesis

TabPFNv2 shows broad-cohort set-level enrichment without a reliable pointwise ranking; TabICLv2 establishes neither property. On the shared 9-dataset × 3-split scope, however, the set-level excess is not established for either readout: TabPFNv2 gives $\Gamma_{\text{flip}} = +1.96$ pp $[-0.93, +5.21]$ $(7/9)$ and TabICLv2 $+1.00$ pp $[-0.21, +2.44]$ $(6/9)$, both Inconclusive (Appendix F.1). The flip–delete contrast remains positive on the shared scope and reaches Support for both readouts ($6/9$ direction consistency for TabPFNv2, $9/9$ for TabICLv2). A complete original-cohort 8×5 robustness check gives the same qualitative pointwise conclusion; Appendix F reports the estimates.

6 Mechanism-Matched Intervention Probe

The supported TabPFNv2 set-level excess (§4) raises a mechanistic follow-up: does reducing attention concentration weaken that excess? The probe divides the L9 query-to-context attention logits by a temperature before the softmax, $A_T = \text{softmax}(z/T)$ for $T \in \{1, 1.25, 1.5, 2\}$, on the same 20-dataset cohort, splits, budget, hub selector and matched-random sets. It changes only that path and never sees which rows are corrupted.

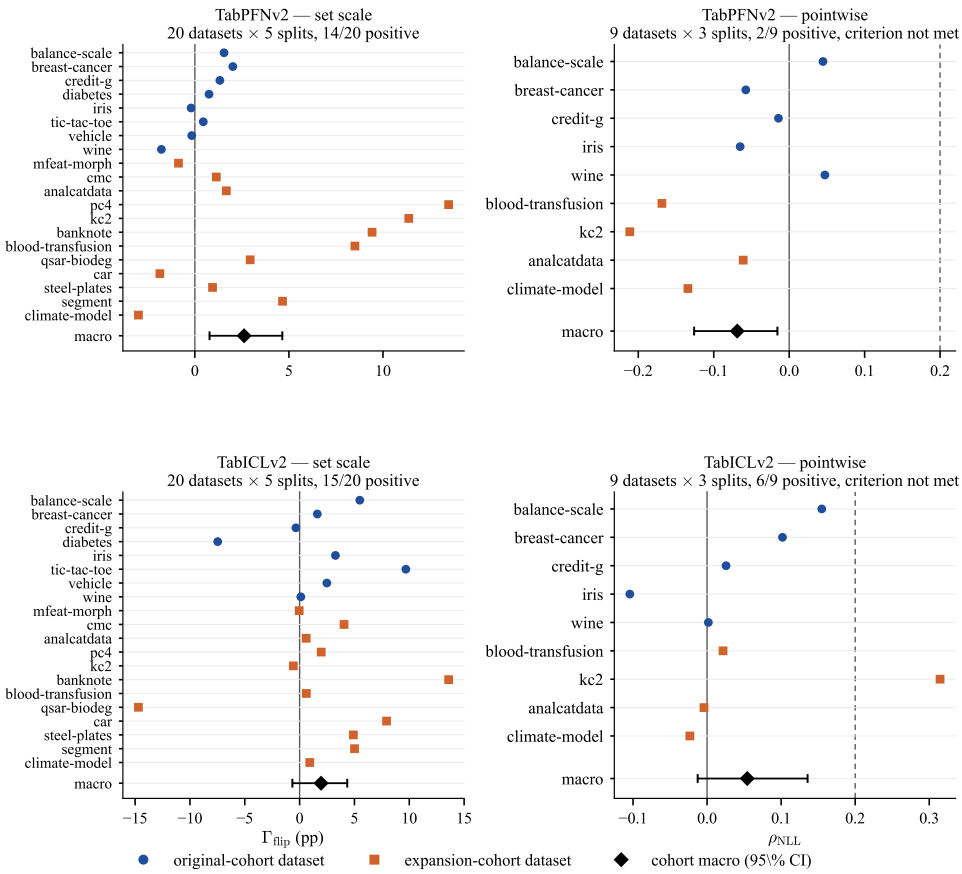


Figure 3: Set-level risk enrichment and pointwise rank correspondence. Left column: the set-level high-vs-matched-random excess Γ_{flip} in accuracy percentage points for each of the 20 datasets at 5 splits. Right column: the pointwise Spearman ρ between each readout's mechanism-aligned query-label-free importance and the single-point damage D_{NLL} across the 9 datasets $\times$ 3 splits of the pointwise study; the dashed line marks the $\rho = 0.20$ reliable-ranker criterion. Within each column one dataset ordering is used; original-cohort datasets are shown as blue circles, expansion-cohort datasets as orange squares, and diamonds mark cohort macro means with dataset-level paired bootstrap 95% CIs.

On the combined cohort the excess falls monotonically in T: at T=1.25, $\Delta\Gamma_{\text{flip}} = -0.19$pp $[-0.31, -0.07]$ with a clean-accuracy change of -0.03pp, and at T=2 it is -0.50pp $[-0.93, -0.16]$. All reductions in Γ_{flip} remain below the 1pp practical-effect scale (Fig. 2B).

Attention concentration appears to contribute to part of the excess, but the sub-1pp reductions show that simple de-concentration accounts for only a limited part of the observed vulnerability. We therefore treat this intervention as a mechanistic probe rather than a mitigation result. Appendix E records the probe grid and the position-oracle case study.

7 Discussion

The attention-vote score shows set-level enrichment on the broader 20×5 Accuracy analysis and weak pointwise correspondence on the 9×3 study. On the shared 9×3 scope, the set-level excess is inconclusive.

Tail selection does not recover useful pointwise enrichment: across the top 10%, 15% and 20% importance fractions neither readout shows a stable positive increase in single-point damage, and for

TabPFNv2 the differences are negative at all three fractions, with the top-20% comparison excluding zero in the negative direction (Appendix F.2). Secondary set-level metrics likewise show strong metric and cohort dependence: for TabPFNv2 the NLL excess is negative on the original cohort, positive on the expansion cohort and positive but inconclusive on the combined cohort, while the true-class-probability excess stays negative across cohorts (Appendix G). The supported broad-cohort result is therefore specific to Accuracy rather than a metric-independent vulnerability claim. Whether multi-row interactions or shared set structure contribute to the set-level effect remains unresolved.

Set-level excess varies across cohorts and readouts: TabPFNv2 is supported on the expansion and combined Accuracy analyses, whereas TabICLv2 reaches Support on no cohort. The evidence is therefore limited to the tested readouts, datasets, and mechanism-specific selectors; for TabICLv2 the negative pointwise result concerns the leave-one-out prototype-drift operationalization, not prototype-based scores that model the flipped destination class.

The tested readout-aligned scores describe importance under the clean context but do not provide reliable pointwise rankings of wrong-label damage in the tested scope. For the tested scores, clean-context prediction importance and pointwise corruption sensitivity are therefore distinct quantities.

Pointwise protection therefore requires direct validation of the score used to prioritize examples. Attention de-concentration moves the supported excess in the expected direction, but the reductions stay below the practical-effect scale at every prespecified temperature, so we treat it as a mechanistic probe rather than a mitigation result.

8 Conclusion and Limitations

Our results separate two properties that are easy to conflate: enrichment of a vulnerable set and ranking of single-point wrong-label damage. For TabPFNv2, the Accuracy-based set-level excess is supported on the combined 20-dataset cohort but is not established on the shared 9-dataset $\times$ 3-split scope, while its own importance score does not meet the reliable-ranker criterion pointwise. TabICLv2 establishes neither stable set-level excess nor reliable pointwise ranking. For both readouts, retaining wrong labels is more harmful on average than deleting the same rows.

Limitations. Twenty dataset-level units still produce wide intervals, so effects near the 1pp practical-effect threshold remain imprecisely estimated, and both the 1pp and $\rho \geq 0.20$ thresholds are study-specific operational criteria. The evidence is limited to eligible OpenML-CC18 classification datasets and the two tested readout settings. The set-level study uses 20 datasets $\times$ 5 splits and the pointwise study 9 datasets $\times$ 3 splits; on the shared 9×3 scope the set-level excess is not established (Appendix F.1). The readout-based scores are operational proxies for mechanisms characterized in prior work, and the only signal meeting the pointwise ranking criterion on the original five datasets uses query true labels. Finally, we study one corruption type at one 15% budget: for multiclass examples the pointwise intervention evaluates one wrong-target realization rather than an expectation over all wrong targets, and the intervention probe modifies a single TabPFNv2 attention path.

AI use statement

We used generative AI tools to assist with manuscript drafting and language editing, and to write and refactor parts of the analysis, table-generation and figure-generation code released with this paper. We did not use generative AI tools to design the studies, to select research questions, or to produce experimental results: every quantity reported in this paper is computed from the frozen artifacts described in Appendix A by the released analysis scripts, and the agreement between the numbers in the text and those artifacts is checked by automated consistency tests that accompany the release. All AI-assisted text, code and references were reviewed and edited by the authors, who take responsibility for the final content of this work.

Ethics statement

This study examines the harm of wrong labels in tabular foundation models using public benchmark datasets and no human subjects. The corruption operator we study is applied offline to released artifacts and is used to measure vulnerability, not to attack deployed systems. We therefore see no specific ethical concern raised by this work beyond the general caution that corruption-based evaluation procedures can be repurposed, which is why the intervention we test is reported as a mechanistic probe rather than a defense.

Reproducibility statement

Model configurations, damage definitions, corruption and selection rules, the 1pp practical-effect threshold, and the aggregation and bootstrap conventions are documented in versioned protocol artifacts (Appendices A and A.2; the provenance record carries no independently verifiable external timestamp). All verdicts reported in this paper use the three-state criterion defined in §3, including an exact two-thirds direction-consistency requirement. The models are evaluated through their official implementations under fixed deterministic settings (fp32, TF32 disabled, deterministic cuDNN, fixed ensemble seeds, split seeds, corruption streams, and bootstrap seeds); splits, corruption rules, selectors and estimators are fixed in code. During review, anonymized analysis code, table- and figure-generation scripts, raw per-unit result files, analysis-plan and provenance records, and automated consistency checks are provided as supplementary material accompanying this submission. These materials reconstruct the reported headline quantities from the supplied results.

References

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 1467–1474. Omnipress, 2012.

Marin Biloš, James T. Wilson, Anderson Schneider, and Yuriy Nevmyvaka. A mechanistic study of tabular foundation models. arXiv:2605.21288, 2026. URL `https://arxiv.org/abs/2605.21288`.

Mohamed Djilani, Thibault Simonetto, Karim Tit, Florian Tambon, Paul Récamier, Salah Ghamizi, Maxime Cordy, and Mike Papadakis. On the robustness of tabular foundation models: Test-time attacks and in-context defenses, 2025. arXiv:2506.02978; accepted at IEEE SaTML 2026.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, New York, 1993. ISBN 978-0-412-04231-7.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2023.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

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.

James Hu and Mahdi Ghelichi. Noise immunity in in-context tabular learning: An empirical robustness analysis of TabPFN's attention mechanisms. arXiv:2604.04868, 2026. URL https://arxiv.org/abs/2604.04868.

Roman Jaeschke, Joel Singer, and Gordon H. Guyatt. Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, 10(4):407–415, 1989.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017.

John K. Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018.

Daniël Lakens, Anne M. Scheel, and Peder M. Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations (ICLR 2022)*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930. Curran Associates, Inc., 2020.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning (ICML 2026)*, 2026. URL https://openreview.net/forum?id=SxsyLjIfWB.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2022.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pp. 11–20, 2019.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2422–2437. Association for Computational Linguistics, 2022.

A Analysis plan and provenance

Original study (TabPFNv2). The supplied analysis-plan artifact records the analysis plan used for the original study, and the file hashes shipped with it identify the archived files by content; the record itself carries no independently verifiable external timestamp. Recorded plan elements: model and checkpoint; the 8 datasets and their 5 stratified splits (OpenML IDs: balance-scale 11, breast_cancer 1510, credit-g 31, diabetes 46921, iris 61, tic-tac-toe 50, vehicle 54, wine 187); hub definition (L9 representations, k=5 cosine centrality, top-15%) and flip rule including RNG keys; matched-random counterfactual structure (R=10, class-stratified, disjoint); practical-effect threshold/ROPE; the measurement blocks and their success criteria; and the verdict vocabulary.

Cross-model extension (TabICL). The TabICL arm uses a separate analysis plan with the same study structure and quantities, together with a TabICL-specific influence definition and surrogate-agreement criterion. The v2 checkpoint meets that criterion (15-cell agreement mean 0.973, min $0.820 \geq 0.80$).

Version robustness across TabICL releases. The three officially released TabICL checkpoints were screened by the same prespecified surrogate-agreement criterion as the v2 model (minimum 15-cell agreement ≥ 0.80 between the prototype readout and the native decoder). All releases are evaluated with the same validation script, which covers checkpoint identity (sha256), package compatibility, the readout layer, prototype construction on post-LN representations, Euclidean versus cosine variants, member-wise computation, ensemble aggregation, class-shuffle correction, context/query indexing, sample-ID mapping, and bitwise reconstruction of the native predictions from the extracted representations. It also resolved an implementation mismatch that is specific to this family: the official prediction path forwards all members of a normalization group as a single batch, and the column-embedding layer reuses the first table of a batch (shuffle optimisation), so member-by-member single-row forwards do not reproduce native predictions on v1/v1.1 while they do on v2. The extraction was therefore corrected to the package-faithful group-batched path; v2 is unaffected (the corrected path matches its native predictions to float precision, and the paper's TabICLv2 results were produced by an equivalent path). With the corrected extraction, v1 meets the agreement criterion (minimum 0.815) and v1.1 does not (minimum 0.755); v1 is therefore reported as a version robustness check for the prototype family, and v1.1 is *excluded from downstream causal analysis by the same criterion* and reported only as a fidelity record (Table 9). All releases are screened with the same prespecified metric and threshold.

Dataset expansion. The twelve-dataset expansion uses outcome-independent eligibility criteria over the public OpenML-CC18 pool, with the same split, cap, budget, flip, matching, estimator, practical-effect threshold and bootstrap contracts as the original cohort; the four expansion datasets used for single-point analysis were selected by an outcome-independent characteristic-based rule: the lowest-dimensional binary dataset (blood-transfusion-service-center), a small-sample high-dimensional binary dataset (kc2), a pure-categorical multiclass dataset (analcatdata_dmft), and a small-sample medium-dimensional binary dataset (climate-model-simulation-crashes), one per stratum. The original cohort, the expansion-only cohort and their pre-specified combination are reported separately throughout (Tables 3 and 8). Two protocol-level corrections are recorded in the expansion materials: two datasets' feature counts were corrected to the verified post-encoding shapes, and the context/query cap truncation was made class-stratified (it never binds for the original eight datasets, whose splits reproduce bit-identically).

Table 2: The 20-dataset cohort. The eight original datasets are the original cohort; the twelve expansion datasets were selected from the OpenML-CC18 suite by objective eligibility criteria fixed before any model outcome existed (Appendix A). D3 marks membership in the single-point cohort, selected by a characteristic-based rule specified before any result.

Dataset	OpenML	n	d	K	Role	D3
balance-scale	11	625	4	3	original	✓
breast_cancer	1510	569	30	2	original	✓
credit-g	31	1000	20	2	original	✓
diabetes	46921	768	8	2	original	
iris	61	150	4	3	original	✓
tic-tac-toe	50	958	9	2	original	
vehicle	54	846	18	4	original	
wine	187	178	13	3	original	✓
mfeat-morphological	18	2000	6	10	expansion	
cmc	23	1473	9	3	expansion	
analcatdata_dmft	469	797	4	6	expansion	✓
pc4	1049	1458	37	2	expansion	
kc2	1063	522	21	2	expansion	✓
banknote-authentication	1462	1372	4	2	expansion	
blood-transfusion-service-center	1464	748	4	2	expansion	✓
qsar-biodeg	1494	1055	41	2	expansion	
car	40975	1728	6	4	expansion	
steel-plates-fault	40982	1941	27	7	expansion	
segment	40984	2310	16	7	expansion	
climate-model-simulation-crashes	40994	540	18	2	expansion	✓

Eligibility thresholds: OpenML-CC18 member; no overlap with the original eight; classification with $2 \leq K \leq 10$; zero missing values; $1 \leq d \leq 48$ post-encoding features; $150 \leq n \leq 2500$; near-duplicate registrations of an already selected task were pruned by a deterministic rule fixed before any result (pc3/pc4 keep pc4; kc1/kc2 keep kc2; pc1 excluded as a same-family task of pc4; the classic diabetes id 37 excluded as a name collision with the original cohort's id 46921). Selection used no model outcome.

Additional sensitivity analyses are reported descriptively and are not used for the primary verdicts: practical-effect-threshold sensitivity, the label-aware gradient baseline, the hub-definition grid, the secondary-metric replay and the matching sensitivity (§G). Release-robustness extension: the TabICL surrogate-agreement screening was applied to each release, and the family-specific extraction mismatch was corrected on architecture/source-code grounds (§C). Follow-up intervention probe: §D documents the attention de-concentration probe.

A.1 Single-point estimator disclosure

We use a pooled estimator: one Spearman correlation per dataset over all its single-point flip records across the three splits, then an unweighted dataset macro mean with dataset-level paired bootstrap 95% CI (seed 20260908). We re-checked the estimator under a per-(dataset, split)-then-mean reading: no verdict, success-criterion state or conclusion changes (max per-dataset $|\Delta\rho| = 0.055$, macro $|\Delta\rho| \leq 0.02$, hub-vs-rest identical), only three sub-threshold direction counts differ by one dataset on near-zero signals.

Implementation disclosure. The initial run of the single-point analysis was re-executed after a code/data-pattern error was found; the paper cites the corrected results. The uniform control U's per-point reseeding defect makes it constant within a cell; U is not read as a candidate signal and is not used to calibrate any noise-floor conclusion.

A.2 Operational protocol details

Inference pipeline (TabPFNv2). fp32 determinism (TF32 off, cuDNN deterministic), 8 ensemble members ($n_estimators$=8, $random_state$=0), official preprocessing including native categorical-column handling; feature shuffling permutes columns only, row order is preserved, so raw sample IDs map one-to-one onto context-row indices. L9 definition: zero-indexed block 9; interventions act on that block's row (cross-sample) attention module (6 heads attending over rows along feature columns); feature attention and other blocks are untouched. The intervention path

matches the no-intervention baseline (max probability difference $\leq 3.6 \times 10^{-7}$) and the deletion identity $C'_{\mathcal{P}} \setminus \mathcal{P} = \mathcal{C} \setminus \mathcal{P}$ holds, so interventions touch only the specified positions. Hub representations: block-9 output cell tokens mean-pooled over the feature-group axis and averaged over the 8 ensemble members.

Flip rule and RNG. Selected labels are flipped uniformly to one of the other $K-1$ classes: $u \sim \text{Uniform}\{0, \ldots, K-2\}$; $y_{\text{wrong}} = u$ if $u < y$ else $u+1$; swap for K=2. Matched-random sets are drawn class-stratified to match the high set's original-class counts; the same uniform-other-class rule is then applied to each set separately. Each position set uses an independent deterministic RNG stream (FNV-1a-keyed PCG64), keyed identically on both model arms. Across the complete stored scan (20 datasets $\times$ 5 splits $\times$ 10 replicates, 1,000 matched-random sets), the class-composition fallback triggered in 50 sets (5%), all in expansion-cohort `steel-plates-fault` and spread evenly across its five splits (10 sets per split, i.e. every replicate of every split); the total corruption budget was preserved in every set (only the exact class composition of those sets deviates), and the original eight-dataset cohort has no triggered sets.

L9 intervention and exposure matching (repair case study). The intervention modifies only query-row-to-context-row attention weights in L9 row attention on the test-row branch, preserving the multi-query head-0 K/V structure (all 6 query heads treated identically); intra-context, feature attention and other blocks are untouched. Hard mask: $A_{qi} = 0,\ i \in \mathcal{P}$, then renormalize over remaining keys; soft: $A_{qi} \leftarrow \alpha A_{qi}$ ($\alpha \in \{0.25, 0.5, 0.75\}$, unnormalized); α=1 with nonempty $\mathcal{P}$ is the no-op validation path. The oracle knows the corruption positions but not the true labels, and uses neither clean activations nor test labels. Exposure (post-corruption forward, pre-intervention): $X(\mathcal{P}) = \frac{1}{Q}\sum_q \sum_{i \in \mathcal{P}} \bar{A}^{\text{poison}}_{qi}$ with $\bar{A}$ averaged over heads and members. The wrong-position control set W satisfies $W \cap \mathcal{P} = \varnothing$, $|W| = |\mathcal{P}|$, built by the recorded algorithm: per-class quotas from the target set's counts; within each class, non-P candidates greedily in ascending per-point-exposure deviation (ties by sample ID); shortfalls filled by exposure proximity; final set in ascending position order. The match succeeds iff $|X(W) - X(P)|/(X(P) + 10^{-12}) \leq 0.10$, otherwise that dataset's exposure match is invalid and its G_{position} receives no position-specific interpretation, where $G_{\text{position}} \triangleq R_{\text{correct}} - R_{\text{wrong}}$.

B Set-level per-dataset results (both models)

Table 5 reports TabPFNv2 per-dataset split means with the macro table in Table 6; Table 7 reports TabICLv2 per-dataset values (means over the 5 splits) with its macro row. Body summary and verdicts: §4.

Table 3: Set-level causal effects by cohort (accuracy pp). Original = the 8 original datasets; expansion-only = the 12 expansion datasets; combined = the pre-specified 20-dataset synthesis. Dataset-level bootstrap 95% CI (seed 20260908, 10,000 resamples); direction counts in parentheses; panel (b) continues the same rows.

Cohort	Model	n	J_{hi}	Γ_{fl}	Γ_{del}
Original	TabPFNv2	8	$3.43[2.06, 5.00](8/8)$	$0.49[-0.33, 1.23](5/8)$	$0.01[-0.16, 0.22](3/8)$
	TabICLv2	8	$4.08[2.30, 6.19](8/8)$	$1.86[-1.46, 4.99](6/8)$	$0.37[-0.21, 1.03](4/8)$
Expansion-only	TabPFNv2	12	$5.94[3.00, 9.00](10/12)$	$4.03[1.19, 7.12](9/12)$	$0.08[-0.75, 1.03](3/12)$
	TabICLv2	12	$4.28[2.44, 6.37](12/12)$	$2.02[-1.82, 5.31](9/12)$	$0.32[-0.63, 1.23](8/12)$
Combined	TabPFNv2	20	$4.93[3.04, 6.90](18/20)$	$2.62[0.78, 4.65](14/20)$	$0.05[-0.47, 0.62](6/20)$
	TabICLv2	20	$4.20[2.87, 5.74](20/20)$	$1.96[-0.66, 4.34](15/20)$	$0.34[-0.28, 0.93](12/20)$

Cohort	Model	n	Γ_J	$D^{\text{hi}}_{\text{fl}}$	$D^{\text{hi}}_{\text{del}}$
Original	TabPFNv2	8	$0.49[-0.22, 1.13](6/8)$	$3.81[2.20, 5.47](8/8)$	$0.39[-0.03, 0.84](6/8)$
	TabICLv2	8	$1.49[-1.52, 4.37](6/8)$	$4.77[2.73, 7.10](8/8)$	$0.69[0.06, 1.34](6/8)$
Expansion-only	TabPFNv2	12	$3.95[1.16, 7.04](8/12)$	$6.41[3.40, 9.56](10/12)$	$0.48[-0.11, 1.16](6/12)$
	TabICLv2	12	$1.71[-2.35, 4.91](10/12)$	$5.17[3.20, 7.33](11/12)$	$0.89[0.27, 1.58](9/12)$
Combined	TabPFNv2	20	$2.57[0.78, 4.56](14/20)$	$5.37[3.40, 7.39](18/20)$	$0.44[0.05, 0.88](12/20)$
	TabICLv2	20	$1.62[-1.04, 3.94](16/20)$	$5.01[3.56, 6.61](19/20)$	$0.81[0.36, 1.29](15/20)$

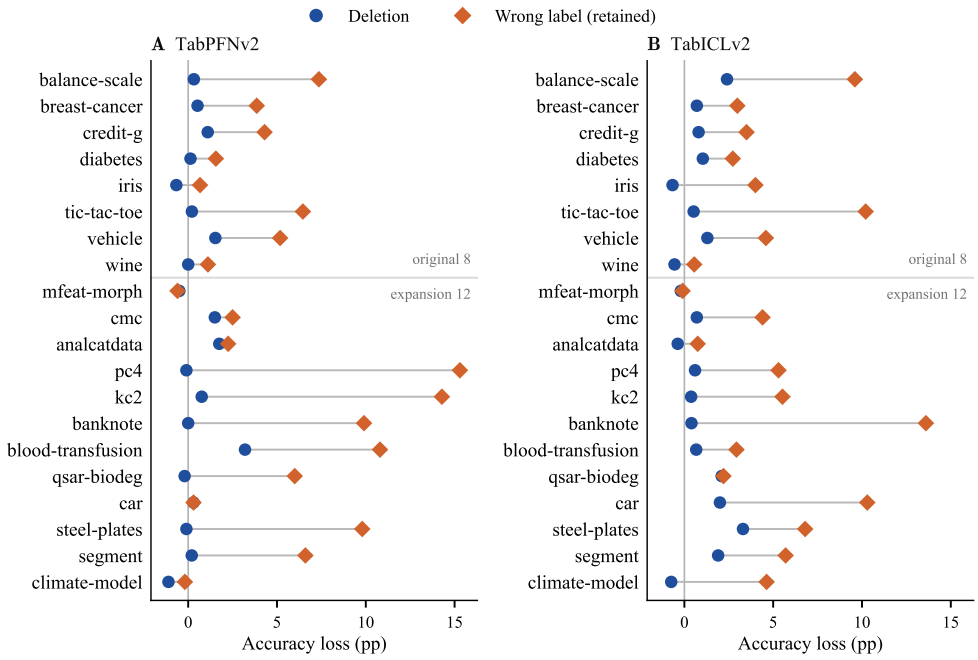


Figure 4: Per-dataset wrong-label versus deletion damage. Accuracy loss from retaining wrong labels and from deleting the same mechanism-defined high-influence rows, shown separately for all 20 datasets. The left panel shows TabPFNv2 and the right panel shows TabICLv2. Points connected on the same row refer to the same dataset; a horizontal divider separates the original eight datasets from the twelve-dataset expansion. The flip–delete difference is positive on 18/20 datasets for TabPFNv2 and 20/20 for TabICLv2. Dataset-level values correspond to the set-level results reported in the appendix tables.

Table 4: Set-level per-dataset evidence (accuracy pp): J_{hi} and Γ_{flip} for both models on all 20 datasets. Original-cohort rows reproduce the original values and the original macro row cites the stored archive; expansion and combined quantities use the same estimator. Macro rows: dataset-level paired bootstrap 95% CI and direction counts.

Dataset	Cohort	TabPFNv2		TabICLv2	
		J_{hi}	Γ_{flip}	J_{hi}	Γ_{flip}
balance-scale	orig.	+7.04	+1.55	+7.20	+5.49
breast_cancer	orig.	+3.33	+2.02	+2.28	+1.61
credit-g	orig.	+3.20	+1.33	+2.70	−0.34
diabetes	orig.	+1.43	+0.75	+1.69	−7.47
iris	orig.	+1.33	−0.20	+4.67	+3.27
tic-tac-toe	orig.	+6.25	+0.45	+9.69	+9.69
vehicle	orig.	+3.65	−0.16	+3.29	+2.48
wine	orig.	+1.11	−1.78	+1.11	+0.11
mfeat-morphological	exp.	−0.10	−0.87	+0.10	−0.04
cmc	exp.	+1.00	+1.14	+3.70	+4.05
analcatdata_dmft	exp.	+0.50	+1.67	+1.12	+0.60
pc4	exp.	+15.40	+13.49	+4.70	+1.97
kc2	exp.	+13.52	+11.37	+5.14	−0.57
banknote-authentication	exp.	+9.90	+9.42	+13.20	+13.59
blood-transfusion-service-center	exp.	+7.60	+8.51	+2.27	+0.61
qsar-biodeg	exp.	+6.20	+2.94	+0.10	−14.70
car	exp.	+0.00	−1.86	+8.30	+7.93
steel-plates-fault	exp.	+9.90	+0.94	+3.50	+4.90
segment	exp.	+6.40	+4.66	+3.80	+5.01
climate-model-simulation-crashes	exp.	+0.93	−3.00	+5.37	+0.93
Original 8	macro	+3.43 (8/8)	+0.49 (5/8)	+4.08 (8/8)	+1.86 (6/8)
Expansion 12	macro	+5.94 (10/12)	+4.03 (9/12)	+4.28 (12/12)	+2.02 (9/12)
Combined 20	macro	+4.93 (18/20)	+2.62 (14/20)	+4.20 (20/20)	+1.96 (15/20)

Table 5: D1 main results (Accuracy; per-dataset split means). $\Gamma_{\text{flip}} = D_{\text{flip}}(P_{\text{hub}}) - D_{\text{flip}}(P^{(r)}_{\text{rand}})$; Γ_{delete} and Γ_J are isomorphic.

Dataset	Clean	Hub Flip	Rand Flip	Hub Del	Rand Del	$J(P_{hub})$	$J(P_{rand})$	Γ_{flip}	Γ_{del}	Γ_J
balance-scale	0.974	0.901	0.916	0.971	0.971	0.070	0.054	0.016	0.000	0.016
breast_cancer	0.970	0.932	0.952	0.965	0.967	0.033	0.015	0.020	0.002	0.018
credit-g	0.779	0.736	0.749	0.768	0.774	0.032	0.025	0.013	0.006	0.007
diabetes	0.788	0.773	0.780	0.787	0.788	0.014	0.007	0.008	0.001	0.007
iris	0.967	0.960	0.958	0.973	0.972	0.013	0.014	-0.002	-0.001	-0.001
tic-tac-toe	0.996	0.931	0.936	0.994	0.993	0.062	0.057	0.004	-0.001	0.005
vehicle	0.860	0.808	0.807	0.845	0.841	0.036	0.035	-0.002	-0.003	0.002
wine	1.000	0.989	0.971	1.000	0.998	0.011	0.027	-0.018	-0.002	-0.016

Table 6: D1 macro means (unweighted dataset macro average; dataset-level paired bootstrap 95% CI; 1pp practical-effect threshold).

Quantity	mean	95% CI	pos. datasets
$J(P_{\text{hub}})$ (net harm vs. deletion)	3.43	[2.06, 5.00]	8/8
$J(P^{(r)}_{\text{rand}})$ (matched-random net harm)	2.94	[1.80, 4.18]	8/8
$D_{\text{flip}}(P_{\text{hub}})$ (hub flip damage)	3.81	[2.20, 5.47]	8/8
$D_{\text{flip}}(P^{(r)}_{\text{rand}})$ (matched-random flip damage)	3.32	[1.97, 4.77]	8/8
$D_{\text{delete}}(P_{\text{hub}})$ (hub deletion loss)	0.39	[-0.03, 0.84]	6/8
Γ_{flip}	0.49	[-0.33, 1.23]	5/8
Γ_{delete}	0.01	[-0.16, 0.22]	3/8
Γ_J	0.49	[-0.22, 1.13]	6/8

Table 7: TabICLv2 cross-model D1 results (per-dataset mean over 5 splits of accuracy, pp); J = net flip harm vs. deletion; $\Gamma_{\text{flip}}/\Gamma_{\text{delete}}/\Gamma_J$ = contrasts vs. matched-random (definitions and protocol provided in the supplementary material).

Dataset	J (pp)	Γ_{flip} (pp)	Γ_{delete} (pp)	G_J (pp)
balance-scale	7.20	5.49	2.16	3.33
breast_cancer	2.28	1.61	0.74	0.88
credit-g	2.70	-0.34	-0.21	-0.13
diabetes	1.69	-7.47	-0.49	-6.97
iris	4.67	3.27	-0.33	3.60
tic-tac-toe	9.69	9.69	0.45	9.24
vehicle	3.29	2.48	1.19	1.29
wine	1.11	0.11	-0.56	0.67
Macro (95% CI)	4.08 [2.30, 6.19]	1.86 [-1.46, 4.99]	0.37 [-0.21, 1.03]	1.49 [-1.52, 4.37]

C Single-point per-dataset results (both models)

Table 10 reports the TabPFNv2 signals' predictability per damage measure with evidence-shape labels on the original cohort. The oracle gradient baseline I^{grad} (Eq. 9) is reported as a macro row in Table 12 and in full in Table 11; Table 13 reports the per-dataset ρ of the primary damage measure for both models and for the eligible TabICL v1 release on the combined 9-dataset cohort. Body summary and verdicts: §5.

Table 8: Single-point importance–vulnerability rank correlation by cohort (Spearman ρ, pooled over splits within each dataset, macro over datasets). The success bar is $\rho \geq 0.20$ with CI lower bound above zero and at least two thirds of the datasets agreeing in direction.

Cohort	Signal	ρ_{NLL}
Original 5	TabPFNv2 H	$-0.009[-0.052, +0.034](2/5)$
	TabPFNv2 M	$-0.062[-0.100, -0.030](0/5)$
	TabPFNv2 $Ebar$	$-0.082[-0.174, +0.014](1/5)$
	TabICL I^{proto}	$+0.036[-0.042, +0.108](4/5)$
Expansion-only 4	TabPFNv2 H	$-0.144[-0.192, -0.088](0/4)$
	TabPFNv2 M	$+0.113[+0.029, +0.214](4/4)$
	TabPFNv2 $Ebar$	$+0.082[+0.007, +0.227](4/4)$
	TabICL I^{proto}	$+0.077[-0.014, +0.235](2/4)$
Combined 9	TabPFNv2 H	$-0.069[-0.126, -0.016](2/9)$
	TabPFNv2 M	$+0.015[-0.052, +0.095](4/9)$
	TabPFNv2 $Ebar$	$-0.009[-0.096, +0.089](5/9)$
	TabICL I^{proto}	$+0.054[-0.013, +0.136](6/9)$

Cohort	Signal	ρ_{prob}	ρ_{Acc}
Original 5	TabPFNv2 H	$-0.017[-0.056, +0.024](2/5)$	$-0.004[-0.040, +0.034](3/5)$
	TabPFNv2 M	$-0.173[-0.246, -0.125](0/5)$	$+0.045[+0.005, +0.092](4/5)$
	TabPFNv2 $Ebar$	$-0.218[-0.305, -0.105](1/5)$	$+0.026[-0.015, +0.068](3/5)$
	TabICL I^{proto}	$-0.028[-0.131, +0.097](1/5)$	$+0.022[-0.014, +0.053](3/5)$
Expansion-only 4	TabPFNv2 H	$+0.054[-0.106, +0.173](3/4)$	$+0.030[-0.044, +0.103](2/4)$
	TabPFNv2 M	$-0.150[-0.211, -0.053](0/4)$	$+0.029[-0.060, +0.140](2/4)$
	TabPFNv2 $Ebar$	$-0.148[-0.417, +0.051](1/4)$	$+0.026[+0.008, +0.055](4/4)$
	TabICL I^{proto}	$+0.062[-0.019, +0.177](3/4)$	$-0.007[-0.093, +0.063](2/4)$
Combined 9	TabPFNv2 H	$+0.015[-0.055, +0.084](5/9)$	$+0.011[-0.029, +0.055](5/9)$
	TabPFNv2 M	$-0.163[-0.214, -0.109](0/9)$	$+0.038[-0.013, +0.093](6/9)$
	TabPFNv2 $Ebar$	$-0.187[-0.309, -0.070](2/9)$	$+0.026[-0.001, +0.052](7/9)$
	TabICL I^{proto}	$+0.012[-0.068, +0.097](4/9)$	$+0.009[-0.037, +0.046](5/9)$

Table 9: TabICL release fidelity and version-robustness (one prototype-readout family, not independent model families). Fidelity is the surrogate agreement of the prototype readout against the native decoder; the prespecified criterion is min ≥ 0.80. ρ columns are macro Spearman over the shared D3 cohort; a release excluded by the fidelity criterion has no causal row.

Release	Fid. mean	Fid. min	Eligible	ρ_{NLL}	Verdict
v1	0.953	0.815	yes	$+0.088[-0.048, +0.227](6/9)$	criterion not met
v1.1	0.938	0.755	no	–	fidelity exclusion
v2	0.973	0.820	yes	$+0.054[-0.013, +0.136](6/9)$	criterion not met

A release whose fidelity minimum falls below the prespecified criterion min ≥ 0.80 is *excluded from downstream causal analysis by the same prespecified criterion* and is reported only as a fidelity record; both eligible releases remain below the reliable-ranker criterion.

Table 10: Single-point signal predictability on the original 5-dataset cohort: dataset macro means and 95% CI of per-dataset Spearman ρ (dataset-level paired bootstrap; success criterion: a signal is a positive predictor iff $\rho \geq 0.20 \wedge \mathrm{CI}_{\mathrm{lower}} > 0 \wedge$ at least two thirds of the datasets direction-consistent. Row-level shape labels (near-zero / negative / not-met) summarize the interval pattern: negative = CI entirely below zero, near-zero = CI covering zero with $|\rho| \leq 0.03$, and not-met = otherwise).

Signal	Damage	ρ	95% CI	pos. datasets	label
H (hubness)	D^{NLL}	-0.009	[-0.052, 0.034]	2/5	near-zero
M (max attention)	D^{NLL}	-0.062	[-0.100, -0.030]	0/5	negative
$\bar{E}$ (mean exposure)	D^{NLL}	-0.082	[-0.174, 0.014]	1/5	not-met
H (hubness)	D^{prob}	-0.017	[-0.056, 0.024]	2/5	near-zero
M (max attention)	D^{prob}	-0.173	[-0.246, -0.125]	0/5	negative
$\bar{E}$ (mean exposure)	D^{prob}	-0.218	[-0.305, -0.105]	1/5	negative
H (hubness)	D^{Acc}	-0.004	[-0.040, 0.034]	3/5	near-zero
M (max attention)	D^{Acc}	0.045	[0.005, 0.092]	4/5	not-met
$\bar{E}$ (mean exposure)	D^{Acc}	0.026	[-0.015, 0.068]	3/5	near-zero

Table 11: Original 5-dataset cohort: every tested signal against all three single-point damage measures (macro Spearman ρ, 95% CI, direction count). The oracle row is query-label-aware and offline; all other rows are query-label-free and pre-corruption.

Signal	D^{NLL}	D^{prob}	D^{Acc}
H (hubness)	$-0.009\,[-0.052, +0.034]\,(2/5)$	$-0.017\,[-0.056, +0.024]\,(2/5)$	$-0.004\,[-0.040, +0.034]\,(3/5)$
M (max attention)	$-0.062\,[-0.100, -0.030]\,(0/5)$	$-0.173\,[-0.246, -0.125]\,(0/5)$	$+0.045\,[+0.005, +0.092]\,(4/5)$
$\bar{E}$ (mean exposure)	$-0.082\,[-0.174, +0.014]\,(1/5)$	$-0.218\,[-0.305, -0.105]\,(1/5)$	$+0.026\,[-0.015, +0.068]\,(3/5)$
TabICLv2 I^{proto}	$+0.036\,[-0.042, +0.108]\,(4/5)$	$-0.028\,[-0.131, +0.097]\,(1/5)$	$+0.022\,[-0.014, +0.053]\,(3/5)$
I^{grad} (oracle)	**+0.387** $[+0.254, +0.541]\,(5/5)$	**+0.231** $[+0.152, +0.293]\,(5/5)$	$+0.226\,[+0.167, +0.290]\,(5/5)$

Note: under the criterion ($\rho \geq 0.20$, CI lower bound > 0, at least two thirds of the datasets) no query-label-free signal passes; I^{grad} passes on all three damage measures.

C.1 Signal definitions

TabICL prototype drift. Within ensemble member m, with post-LN representation rows $R^{(m)}$ and class prototypes $c_k^{(m)}$ (Euclidean means of the context rows of class k, class-shuffle corrected), the leave-one-out drift of point i of class y_i is

$$I_i^{(m)} = \tfrac{1}{Q} \sum_q \left[\|R_q^{(m)} - c_{y_i}^{(m,-i)}\|_2 - \|R_q^{(m)} - c_{y_i}^{(m)}\|_2 \right], \qquad I_i^{\mathrm{TabICL}} = \tfrac{1}{M} \sum_m I_i^{(m)}, \tag{8}$$

where $c_k^{(m,-i)}$ is the prototype recomputed without i ($I_i^{(m)} = 0$ if class y_i has a single member). The class-prototype readout itself is established by Biloš et al. (2026); the leave-one-out drift is our operationalization of it in this study and is not an importance method released with TabICL. This result concerns the pre-corruption leave-one-out source-prototype drift used here; it does not characterize all possible prototype-based pointwise risk scores, including scores that explicitly model the flipped destination class.

Gradient oracle. Under the verified L9 label-vote readout $s_q(y) = \sum_i A_{qi} s_i(y)$ (Biloš et al., 2026), the query loss $\ell_q = -\log P_q(y_q)$ has first-order sensitivity $\partial \ell_q / \partial s_i(y) = A_{qi}(P_q(y) - \mathbf{1}\{y{=}y_q\})$ with respect to point i's label vote, and the expected first-order loss change of a uniform flip to the other $K-1$ classes, averaged over queries, defines

$$\begin{aligned} I_i^{\mathrm{grad}} &= \tfrac{1}{Q} \sum_q A_{qi}\, c_q(i), \\ c_q(i) &= \tfrac{1 - P_q(y_i)}{K-1} - \tfrac{\mathbf{1}\{y_q \neq y_i\}}{K-1} - P_q(y_i) + \mathbf{1}\{y_i = y_q\}. \end{aligned} \tag{9}$$

The indicator terms use the query true labels y_q, so I^{grad} is a query-label-aware offline oracle diagnostic that is not computable at deployment and is not a detector. This derivative is exact for the softmax label-vote readout when the attention weights are held fixed; in the reported diagnostic P_q is evaluated with the native model's stored clean predictive probabilities, so I^{grad} is a first-order readout-aligned diagnostic rather than the exact gradient of the full native network.

Table 12: Importance–vulnerability results on the combined 9-dataset cohort under the primary damage measure D^{NLL}: macro Spearman ρ of every tested signal against single-point wrong-label damage (95% CI; direction count). Non-oracle rows are query-label-free and pre-corruption; the gradient row is the query-label-aware offline oracle of Eq. 9.

Signal (query-label-free)	ρ_{NLL}	Reading
TabPFNv2 H (hubness)	-0.069 [-0.126, -0.016] (2/9)	criterion not met
TabPFNv2 M (max attention)	$+0.015$ [-0.052, $+0.095$] (4/9)	near-zero
TabPFNv2 $\bar{E}$ (mean exposure)	-0.009 [-0.096, $+0.089$] (5/9)	near-zero
TabICLv2 I^{proto} (prototype drift)	$+0.054$ [-0.013, $+0.136$] (6/9)	near-zero
I^{grad} (oracle; original 5)	**+0.387** [$+0.254$, $+0.541$] (5/5)	passes

Table 13: Single-point evidence, combined 9-dataset cohort: per-dataset Spearman ρ of the primary D^{NLL} damage with each model's query-label-free mechanism-aligned importance (pooled over the 3 splits); the TabICL v1 column is the eligible-release version check. Macro rows use the dataset-level paired bootstrap.

Dataset	Cohort	TabPFNv2 H	TabICLv2 I^{proto}	TabICL v1
balance-scale	orig.	$+0.045$	$+0.155$	$+0.325$
breast_cancer	orig.	-0.057	$+0.102$	$+0.045$
credit-g	orig.	-0.014	$+0.026$	$+0.077$
iris	orig.	-0.065	-0.104	-0.153
wine	orig.	$+0.047$	$+0.002$	-0.231
blood-transfusion-service-center	exp.	-0.169	$+0.022$	$+0.277$
kc2	exp.	-0.211	$+0.315$	$+0.437$
analcatdata_dmft	exp.	-0.061	-0.004	$+0.089$
climate-model-simulation-crashes	exp.	-0.134	-0.023	-0.079
Original 5	macro	-0.009 (2/5)	$+0.036$ (4/5)	$+0.013$ (3/5)
Expansion 4	macro	-0.144 (0/4)	$+0.077$ (2/4)	$+0.181$ (3/4)
Combined 9	macro	-0.069 (2/9)	$+0.054$ (6/9)	$+0.088$ (6/9)

Figure 5 shows the pointwise importance–damage relationships underlying the dataset-level rank correlations.

D Attention de-concentration probe details

The probe uses the fixed temperature grid $T \in \{1.0, 1.25, 1.5, 2.0\}$, the L9 query-to-context intervention $A_T = \mathrm{softmax}(z/T)$, the 20-dataset cohort with its splits, budget, hub selector and matched-random sets, the three-tier criterion and the stop rule. 400 cells (20 datasets $\times$ 5 splits $\times$ 4 temperatures) were run; the T=1.0 row reproduces all 1,200 stored native cell values exactly (0 mismatches), so the comparison is between the same model with and without the intervention. Under the protocol's three-tier rule all three non-native temperatures land in the *weak positive* tier on the combined cohort: the excess change $\Delta\Gamma_{\mathrm{flip}}$ has a CI upper bound below zero and at least two thirds of the datasets move in the same direction, the clean-accuracy cost stays below 0.5pp, but the reduction never reaches the 1pp practical scale and the high-position damage falls by less than 1pp. Per the protocol this tier is reported as suggestive and the experiment stops; no temperature outside the grid and no per-head, per-dataset or multi-layer variant was tried.

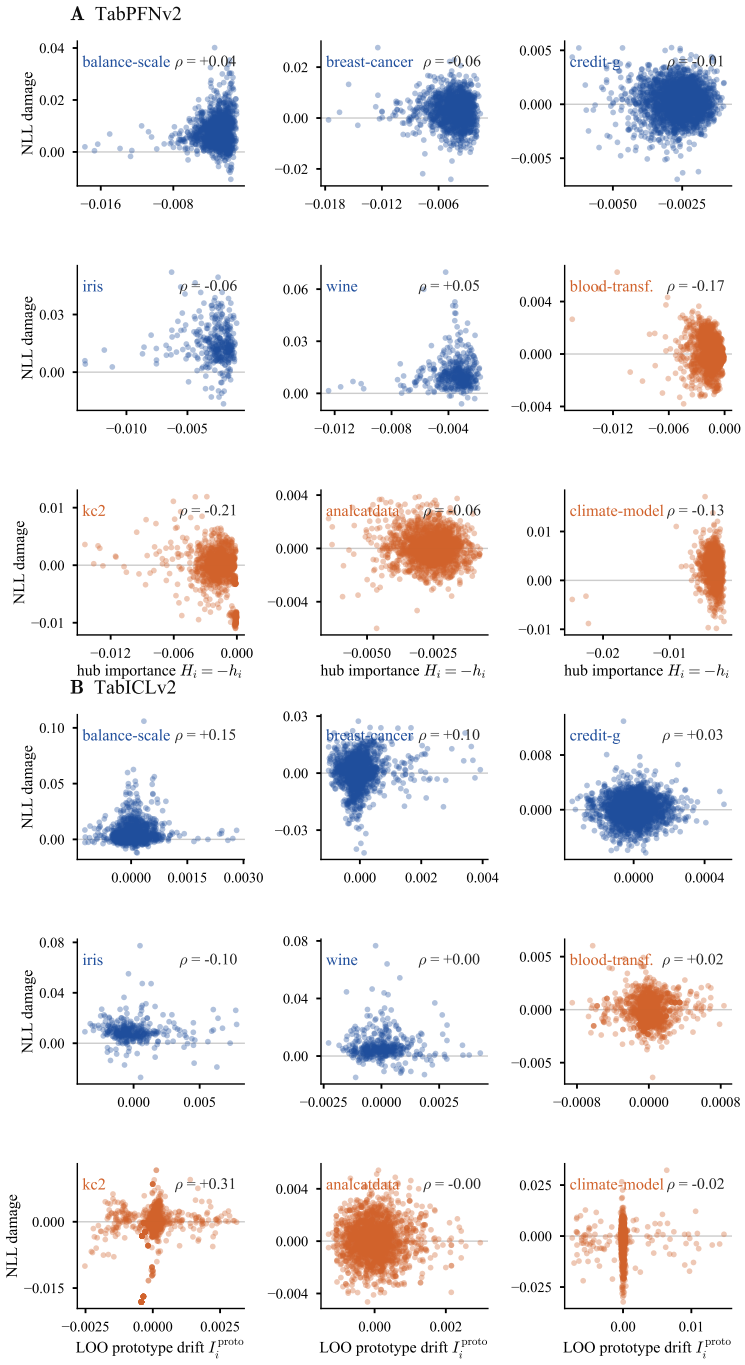


Figure 5: Pointwise importance–damage correspondence across datasets. Small multiples show the relationship between each readout's tested pre-corruption importance signals and single-point NLL damage across the nine datasets in the pointwise study. Each dataset panel pools the single-point interventions from its three splits for visualization; each panel reports the corresponding Spearman ρ. The upper block shows TabPFNv2 hubness (horizontal axis: the oriented hub importance $H_i = -h_i$, so larger values denote more central positions) and the lower block shows TabICLv2 prototype drift. Blue panel names and points denote original-cohort datasets and orange ones denote expansion-cohort datasets. These point-level displays are descriptive; the dataset remains the statistical unit for the paper's macro summaries and confidence intervals.

Table 14: Attention de-concentration (L9 temperature smoothing) by cohort and temperature, accuracy pp. Γ^{T}_{flip} is the high-vs-matched-random excess under temperature T; $\Delta\Gamma$ is its paired change from the native model (T=1); Δclean and the two damage changes are paired dataset-level differences. 95% CI from the dataset-level paired bootstrap (seed 20260908, 10,000 resamples); direction counts in parentheses. The intervention changes only the L9 query-to-context attention.

Cohort	T	Γ^{T}_{flip}	$\Delta\Gamma$ (95% CI; reduced)	Δclean	ΔD^{hi}	ΔD^{rand}
Original 8	1.0	$+0.49$ (5/8)	–	–	–	–
Original 8	1.25	$+0.22$ (5/8)	-0.27 [-0.39, -0.15] (8/8)	-0.20	-0.39	-0.12
Original 8	1.5	$+0.18$ (4/8)	-0.31 [-0.53, -0.12] (7/8)	-0.28	-0.43	-0.12
Original 8	2	$+0.07$ (4/8)	-0.42 [-0.59, -0.25] (8/8)	-0.33	-0.39	$+0.03$
Expansion 12	1.0	$+4.03$ (9/12)	–	–	–	–
Expansion 12	1.25	$+3.90$ (9/12)	-0.13 [-0.33, $+0.04$] (8/12)	$+0.09$	-0.02	$+0.11$
Expansion 12	1.5	$+3.72$ (9/12)	-0.32 [-0.65, -0.04] (9/12)	-0.05	-0.29	$+0.03$
Expansion 12	2	$+3.48$ (9/12)	-0.56 [-1.28, $+0.01$] (7/12)	-0.11	-0.47	$+0.09$
Combined 20	1.0	$+2.62$ (14/20)	–	–	–	–
Combined 20	1.25	$+2.43$ (14/20)	-0.19 [-0.31, -0.07] (16/20)	-0.03	-0.17	$+0.02$
Combined 20	1.5	$+2.30$ (13/20)	-0.32 [-0.53, -0.13] (16/20)	-0.14	-0.35	-0.03
Combined 20	2	$+2.12$ (13/20)	-0.50 [-0.93, -0.16] (15/20)	-0.20	-0.44	$+0.07$

Table 15: Per-dataset change in the high-vs-random excess ($\Delta\Gamma_{\text{flip}}$, pp) under each prespecified temperature; negative means the excess shrank.

Dataset	T=1.25	T=1.5	T=2.0
balance-scale	-0.37	-0.50	-0.50
breast_cancer	-0.37	-0.33	-0.53
credit-g	$+0.00$	-0.54	-0.86
diabetes	-0.30	-0.06	-0.60
iris	-0.07	-0.07	$+0.00$
tic-tac-toe	-0.42	-0.89	-0.47
vehicle	-0.54	$+0.05$	-0.27
wine	-0.11	-0.17	-0.17
mfeat-morphological	-0.20	-0.33	-0.35
cmc	$+0.10$	-0.15	-0.34
analcatdata_dmft	$+0.26$	-0.01	$+0.65$
pc4	$+0.00$	-0.70	-1.48
kc2	-0.17	-0.38	-0.44
banknote-authentication	-0.88	-1.78	-3.65
blood-transfusion-service-center	$+0.32$	$+0.45$	$+0.29$
qsar-biodeg	-0.29	-0.57	-0.92
car	$+0.13$	$+0.28$	$+0.25$
steel-plates-fault	-0.14	-0.10	$+0.16$
segment	-0.19	-0.53	-1.26
climate-model-simulation-crashes	-0.54	$+0.02$	$+0.41$

Figure 6 shows the cohort- and dataset-level structure behind the attention de-concentration probe.

E Repair case study data (TabPFNv2)

Operational definitions of the L9 intervention and the exposure-matched wrong-position control are in §A.2. Table 16 gives per-dataset repair results (primary = hard mask); Table 17 the per-split exposure ratios.

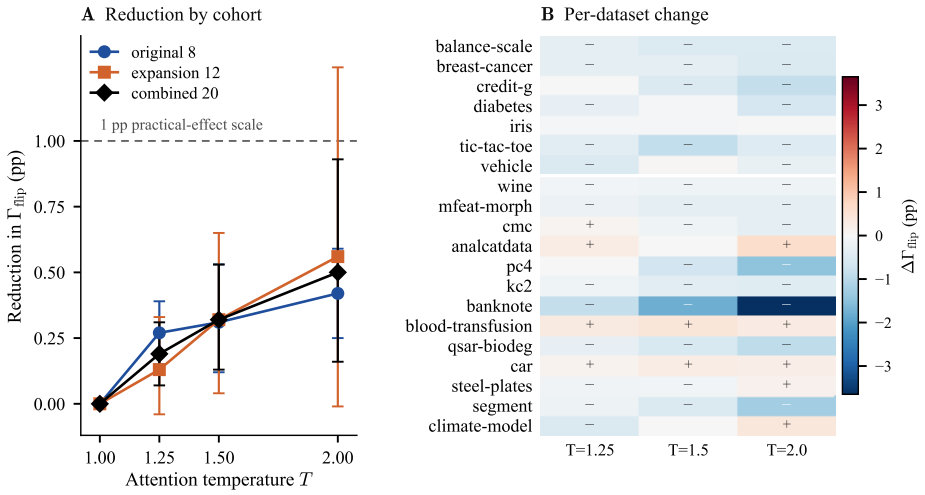


Figure 6: Attention de-concentration across cohorts and datasets. (A) Reduction in the TabPFNv2 high-versus-matched-random excess, $-\Delta\Gamma_{\text{flip}}$, as the L9 attention temperature increases from the native $T = 1$ setting. Curves show the original, expansion, and combined cohorts; error bars are dataset-level bootstrap 95% confidence intervals for the paired change from $T = 1$. The combined cohort contains the original and expansion datasets and is therefore not an independent third cohort. (B) Per-dataset changes in the excess across the three non-native temperatures. The heatmap reports the underlying $\Delta\Gamma_{\text{flip}}$ values, so negative values indicate a reduction in the high-versus-matched-random excess. The display is diagnostic: it shows the heterogeneity underlying the modest combined-cohort reductions rather than establishing a mitigation result.

Table 16: D2 main results (primary = hard mask; Poison = hub-flip accuracy; Match Valid = the dataset's minimum over its 5 splits of the exposure ratio $|X(W) - X(P)|/(X(P) + 10^{-12})$ is ≤ 0.10 (any-split semantics; per-split values in appendix Table 17)). `Match Valid` indicates only that at least one split of the dataset satisfies the exposure-matching threshold; the five-split means in this table are descriptive and are not estimates restricted to exposure-valid split units (restricted summary in Table 32).

Dataset	Poison	Correct L9(hard)	Wrong L9(hard)	Clean Same	R_{correct}	R_{wrong}	G_{position}	Match Valid
balance-scale	0.901	0.904	0.899	0.970	0.003	-0.002	0.005	True
breast_cancer	0.932	0.939	0.933	0.970	0.007	0.002	0.005	False
credit-g	0.736	0.744	0.734	0.777	0.008	-0.002	0.010	False
diabetes	0.773	0.778	0.774	0.790	0.005	0.001	0.004	False
iris	0.960	0.960	0.960	0.967	0.000	0.000	0.000	True
tic-tac-toe	0.931	0.951	0.910	0.998	0.020	-0.021	0.041	False
vehicle	0.808	0.813	0.806	0.858	0.005	-0.002	0.007	False
wine	0.989	0.994	0.989	1.000	0.006	0.000	0.006	True

Table 17: Exposure ratios per dataset and split $|X(W) - X(P)|/(X(P) + 10^{-12})$ (≤ 0.10 valid; bold = any-split valid datasets).

Dataset	s0	s1	s2	s3	s4
balance-scale	0.034	0.011	0.040	0.083	0.052
breast_cancer	0.166	0.274	0.205	0.246	0.239
credit-g	0.127	0.450	0.218	0.130	0.477
diabetes	0.521	0.482	0.375	0.314	0.355
iris	0.020	0.046	0.005	0.050	0.121
tic-tac-toe	0.310	0.298	0.362	0.181	0.349
vehicle	0.483	0.451	0.627	0.529	0.575
wine	0.098	0.080	0.069	0.119	0.011

Three datasets are valid under any-split semantics (balance-scale, iris, wine); only balance-scale (n=1) under the stricter all-splits reading. The position-advantage verdict is Inconclusive under both readings (Appendix E).

Because G_{position} in Table 16 is a mean over all five splits, it is not an estimate restricted to exposure-valid units. Table 32 re-aggregates the same stored per-split outcomes over the 13 of 40 split units whose exposure ratio satisfies the threshold: five datasets contribute no valid split, and on the three that do, G_{position} spans -0.83 to $+0.69$pp. The repair reading is therefore descriptive and remains unsupported under either unit definition.

F Pointwise robustness on the original cohort

Robustness re-check at a larger split count. To assess whether the pointwise conclusion is sensitive to the smaller 9-dataset $\times$ 3-split main-study scope, we repeated the pointwise analysis on all eight original-cohort datasets using five splits per dataset. Table 19 reports the resulting estimates, with Table 18 providing the 9×3 main-study reference. The qualitative conclusion is unchanged: neither of the tested query-label-free importance signals meets the reliable-ranker criterion — the hub importance $H = -h$ stays far below it (-0.019 $[-0.068, +0.032]$ on D_{NLL}, $4/8$ direction-consistent, and $+0.012$ on D_{Acc}), and the prototype importance does not reach it either ($+0.038$ $[-0.005, +0.089]$, $6/8$). This re-check is a robustness analysis on the original cohort, not the 9×3 main pointwise study.

Table 18: Original single-point evidence (9 datasets $\times$ 3 splits), unchanged: macro Spearman ρ of every query-label-free signal against the primary damage measure D_{NLL}, with the release-robustness rows of the prototype family.

Signal	Original 5	Expansion 4	Combined 9
TabPFNv2 H	-0.009 $[-0.052, +0.034]$ (2/5)	-0.144 $[-0.192, -0.088]$ (0/4)	-0.069 $[-0.126, -0.016]$ (2/9)
TabPFNv2 M	-0.062 $[-0.100, -0.030]$ (0/5)	$+0.113$ $[+0.029, +0.214]$ (4/4)	$+0.015$ $[-0.052, +0.095]$ (4/9)
TabPFNv2 $\bar{E}$	-0.082 $[-0.174, +0.014]$ (1/5)	$+0.082$ $[+0.007, +0.227]$ (4/4)	-0.009 $[-0.096, +0.089]$ (5/9)
TabICLv2 I^{proto}	$+0.036$ $[-0.042, +0.108]$ (4/5)	$+0.077$ $[-0.014, +0.235]$ (2/4)	$+0.054$ $[-0.013, +0.136]$ (6/9)

Table 19: Robustness re-check at a larger split count (original 8-dataset cohort, 5 splits, both readouts complete). The macro Spearman ρ of each query-label-free mechanism-aligned signal against the primary D_{NLL} damage and against D_{Acc}, with dataset-level paired bootstrap 95% CIs and direction counts; this replays the 9×3 conclusion reported in Table 18 on a different dataset cohort at 5 splits and does not replace it.

Signal	ρ_{NLL}	ρ_{Acc}
TabPFNv2 H	-0.019 $[-0.068, +0.032]$ (4/8)	$+0.012$ $[-0.014, +0.040]$ (5/8)
TabPFNv2 M	$+0.008$ $[-0.062, +0.095]$ (3/8)	$+0.051$ $[+0.030, +0.078]$ (8/8)
TabPFNv2 $\bar{E}$	-0.046 $[-0.123, +0.038]$ (2/8)	$+0.018$ $[-0.006, +0.045]$ (4/8)
TabICLv2 I^{proto}	$+0.038$ $[-0.005, +0.089]$ (6/8)	$+0.042$ $[+0.010, +0.092]$ (7/8)

F.1 Matched-scope re-aggregation on the shared pointwise scope

To make the set-level and pointwise analyses directly comparable, we re-applied the set-level estimator to exactly the pointwise scope — the nine pointwise datasets at the pointwise splits 0/1/2, read from the artifacts rather than assumed — using the existing per-split outputs, the same dataset-level paired bootstrap and the same three-state criterion applied only for comparability. This is a post-hoc scope alignment and does not replace the 20×5 main analysis (Table 20). On this shared scope the high-vs-random excess is not established for either readout ($\Gamma_{\text{flip}} = +1.96$ pp $[-0.93, +5.21]$ $(7/9)$ for TabPFNv2 and $+1.00$ pp $[-0.21, +2.44]$ $(6/9)$ for TabICLv2, both Inconclusive), while $D_{\text{flip}}(P_{\text{high}})$ itself is supported for both readouts; the flip–delete contrast $J(P_{\text{high}})$ stays positive with an interval excluding zero for both, reaching Support for both readouts ($6/9$ direction consistency for TabPFNv2 and $9/9$ for TabICLv2).

Table 20: Matched-scope re-aggregation on the shared 9 datasets × 3 splits (post-hoc scope alignment, reported as a robustness analysis). The set-level estimator is re-applied to exactly the pointwise scope (the 9 pointwise datasets at the pointwise splits 0/1/2) from the existing per-split outputs; the 20×5 main analysis is unchanged, and the same three-state criterion is applied only for comparability. Values are accuracy percentage points with dataset-level paired bootstrap 95% CIs and direction counts.

Quantity	TabPFNv2	TabICLv2
$D_{\text{flip}}(P_{\text{high}})$	$+4.54$ [$+1.64$, $+7.73$] (7/9) Support	$+3.65$ [$+2.41$, $+5.36$] (9/9) Support
$D_{\text{delete}}(P_{\text{high}})$	$+0.91$ [-0.07, $+1.97$] (5/9) Inconclusive	$+0.44$ [-0.33, $+1.26$] (5/9) Inconclusive
$J(P_{\text{high}})$	$+3.64$ [$+1.15$, $+6.61$] (6/9) Support	$+3.22$ [$+2.01$, $+4.51$] (9/9) Support
Γ_{flip}	$+1.96$ [-0.93, $+5.21$] (7/9) Inconclusive	$+1.00$ [-0.21, $+2.44$] (6/9) Inconclusive

Reading (TabPFNv2, TabICLv2): $D_{\text{flip}}(P_{\text{high}})$: Support, Support; $D_{\text{delete}}(P_{\text{high}})$: Inconclusive, Inconclusive; $J(P_{\text{high}})$: Support, Support; Γ_{flip}: Inconclusive, Inconclusive.

F.2 Tail enrichment of the pointwise scores

The weak pointwise correspondence could in principle hide a usable tail: a score may fail to order all points yet still isolate the most damaging ones. We therefore computed, on the same shared cohort, the mean single-point damage of the top $10/15/20\%$ importance sets against the remaining points and against class-matched random subsets of the same size (fixed grid, no threshold search; Table 21 and Figure 7). Neither readout shows a tail advantage: the tail-minus-rest difference in mean single-point NLL damage is negative at all three fractions for TabPFNv2, with the top-20% interval excluding zero in the negative direction, whereas the TabICLv2 differences are small and their intervals span zero; differences from class-matched random subsets are likewise small and provide no evidence of positive tail enrichment. The decile curves are flat within uncertainty (Figure 7), and the within-tail correlations are below 0.06 in absolute value. This analysis is exploratory and does not add a threshold.

Table 21: Tail enrichment of the pointwise importance scores (exploratory; fixed 10/15/20% grid, no threshold search). Per cell: macro difference in mean single-point NLL damage D_{NLL} between the top-f importance set and the remaining points, with dataset-level paired bootstrap 95% CI and the number of datasets with a positive difference (of 9); the last column contrasts the top-f set with class-matched random subsets of the same size drawn from the non-top points (R=100 draws). Units are 10^{-3} nats throughout: an NLL difference is not on the percentage-point scale, so the 1pp practical-effect threshold does not apply here. TabPFNv2 differences are negative at all three fractions, with the top-20% interval excluding zero in the negative direction; no entry shows positive tail enrichment.

Readout	Top-f	tail − rest (D_{NLL}, 10^{-3} nats)	tail − matched random (D_{NLL}, 10^{-3} nats)
TabPFNv2 ($H = -h$)	10%	-0.61 [-1.61, $+0.15$] (3/9)	-0.71
	15%	-0.63 [-1.49, $+0.04$] (1/9)	-0.58
	20%	-0.67 [-1.47, -0.05] (1/9)	-0.51
TabICLv2 (I^{proto})	10%	$+0.47$ [-0.65, $+1.69$] (6/9)	$+0.06$
	15%	$+0.33$ [-0.89, $+1.60$] (5/9)	$+0.08$
	20%	$+0.55$ [-0.44, $+1.64$] (6/9)	$+0.37$

G Additional sensitivity analyses

The analyses in this section are reported descriptively (§A) and do not enter the primary verdicts. They reuse the same dataset-level estimator (dataset-level paired bootstrap 95% CI, seed 20260908) and are reported only for the original cohort unless stated otherwise. Accuracy with the 1pp practical-effect threshold remains the primary analysis.

G.1 practical-effect-threshold sensitivity

Replaying the three-state criterion on the same stored results (Support requires point estimate $\geq$ practical-effect threshold, CI lower bound > 0, at least two thirds of the datasets direction-consistent;

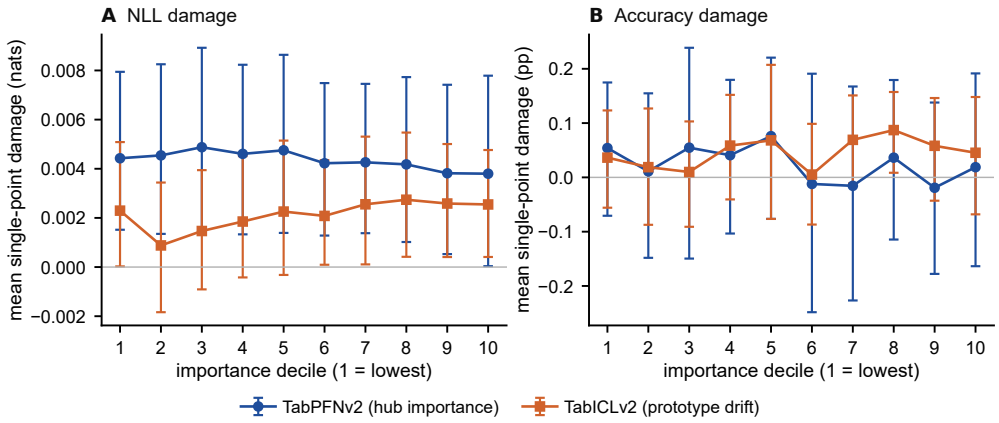


Figure 7: Mean single-point damage by importance decile (exploratory; decile 1 = lowest importance), for the two mechanism-aligned query-label-free importance scores. The two damage measures are not on a common scale, so each has its own panel and its own unit, and the 1pp practical-effect threshold applies to the accuracy panel only. Error bars are dataset-level paired bootstrap 95% CIs over the nine pointwise datasets; both measures stay flat within uncertainty, with no elevated top decile.

ROPE $= \pm 0.5$pp unchanged), the verdicts on the original cohort — the evidence this replay uses — are threshold-insensitive: the position-specific excess quantities (Γ_{flip}, Γ_J) are Inconclusive across the whole 0.5–2.0pp range; even at 0.5pp only R_{correct} and the full-set G_{position} (the latter receiving no position-specific reading, Appendix E) turn Support, and the exposure-valid subset (n=3) stays Inconclusive at every threshold. This replay does not extend to the expansion cohort, where the attention-vote Γ_{flip} verdict is Support at the primary threshold (Table 1(a)).

Table 22: practical-effect-threshold sensitivity: verdicts on the same original-cohort results (mean, CI lower bound, direction count, respectively: Γ_{flip} (0.49pp, -0.33pp, 5/8), Γ_J (0.49pp, -0.22pp, 6/8), R_{correct} (0.67pp, 0.35pp, 7/8), G_{position} full (0.97pp, 0.38pp, 7/8), G_{position} valid subset n=3 (0.35pp, 0.00pp, 2/3)) across practical-effect threshold levels (rule in §G; 1.0pp is the primary threshold, bold column; Γ_{delete}'s ROPE verdict is independent of the practical-effect threshold).

Quantity	0.5pp	**1.0pp**	1.5pp	2.0pp
Γ_{flip}	Inconclusive	Inconclusive	Inconclusive	Inconclusive
Γ_J	Inconclusive	Inconclusive	Inconclusive	Inconclusive
R_{correct}	Support	Inconclusive	Inconclusive	Inconclusive
G_{position} full	Support	Inconclusive	Inconclusive	Inconclusive
G_{position} valid n=3	Inconclusive	Inconclusive	Inconclusive	Inconclusive

G.2 Secondary-metric recomputation (descriptive)

Table 23 recomputes D_{flip}, D_{delete}, J and Γ on the original cohort for NLL, balanced accuracy, macro-F1 and true-class probability; the Accuracy row reports the macro means from §4. Accuracy-type metrics use clean minus poisoned, whereas NLL uses poisoned minus clean. $J(P_{\text{hub}})$ and $J(P_{\text{rand}})$ remain positive with confidence intervals excluding zero on all five metrics. In contrast, no secondary metric provides positive evidence for a high-vs-random excess: Γ_{flip} and Γ_J span zero on balanced accuracy and macro-F1 and have negative point estimates with confidence intervals excluding zero for NLL and true-class probability. The Γ_{delete} intervals cover zero on all metrics. This sensitivity analysis applies only to the original cohort and does not override the expansion/combined Accuracy result.

Table 23: Secondary-metric macro means (dataset-level paired bootstrap 95% CI; Accuracy row as reported in §4; cited in §G). Units: pp (accuracy-type, balanced accuracy, macro-F1), probability (true-class prob.), nats (NLL).

Metric	$D_{\text{flip}}(P_{\text{hub}})$	$D_{\text{flip}}(P^{(r)}_{\text{rand}})$	$J(P_{\text{hub}})$	$J(P^{(r)}_{\text{rand}})$
Accuracy (primary, pp)	3.81 [2.20, 5.47]	3.32 [1.97, 4.77]	3.43 [2.06, 5.00]	2.94 [1.80, 4.18]
Balanced acc (pp)	5.87 [1.14, 13.31]	6.44 [2.16, 13.43]	5.50 [1.05, 12.81]	5.94 [1.99, 12.54]
Macro-F1 (pp)	6.29 [1.82, 13.19]	6.37 [2.24, 12.79]	5.86 [1.69, 12.72]	5.91 [2.11, 12.11]
True-class prob. ($\times 10^{-2}$)	12.85 [9.25, 16.13]	15.61 [11.68, 19.26]	12.06 [8.27, 15.35]	15.03 [11.21, 18.59]
NLL (nats $\times 10^{-2}$)	16.02 [10.53, 20.76]	19.13 [13.19, 24.61]	16.88 [11.43, 21.79]	20.13 [13.97, 25.86]

Metric	Γ_{flip}	Γ_J	Γ_{delete}
Accuracy (primary, pp)	0.49 [−0.33, 1.23]	0.49 [−0.22, 1.13]	0.01 [−0.16, 0.22]
Balanced acc (pp)	−0.57 [−1.49, 0.37]	−0.43 [−1.48, 0.60]	−0.14 [−0.32, 0.06]
Macro-F1 (pp)	−0.08 [−1.00, 0.87]	−0.05 [−1.02, 0.89]	−0.03 [−0.19, 0.16]
True-class prob. ($\times 10^{-2}$)	−2.76 [−4.02, −1.59]	−2.97 [−4.50, −1.50]	0.20 [−0.16, 0.55]
NLL (nats $\times 10^{-2}$)	−3.11 [−4.66, −1.76]	−3.26 [−4.45, −2.15]	0.15 [−0.41, 0.65]

Table 24: Secondary-metric sensitivity across the three cohorts (robustness; dataset-level paired bootstrap 95% CI, direction counts). Accuracy-type metrics are in percentage points, true-class probability in $\times 10^{-2}$ probability units and NLL in $\times 10^{-2}$ nats; accuracy-like damages are clean minus poisoned and NLL is poisoned minus clean, so positive always means more damage for the high-influence set. Original-cohort cells are the sources of Table 23; the expansion and combined columns are the 20×5 recompute. The three cohorts disagree in sign and significance for NLL and true-class probability; no secondary metric enters the primary verdicts.

Metric	Original 8	Expansion 12	Combined 20
High-influence vs matched-random excess, Γ_{flip}			
Accuracy (pp)	+0.49 [−0.33, +1.23] (5/8)	+4.03 [+1.19, +7.12] (9/12)	+2.62 [+0.78, +4.65] (14/20)
Bal. acc (pp)	−0.57 [−1.49, +0.37] (2/8)	+0.84 [−6.24, +6.24] (9/12)	+0.28 [−4.01, +3.56] (11/20)
Macro-F1 (pp)	−0.08 [−1.00, +0.87] (3/8)	+1.88 [−5.73, +7.76] (9/12)	+1.10 [−3.51, +4.70] (12/20)
True-class prob. ($\times 10^{-2}$)	−2.76 [−4.02, −1.59] (0/8)	−0.77 [−2.06, +0.48] (6/12)	−1.57 [−2.60, −0.56] (6/20)
NLL (nats $\times 10^{-2}$)	−3.11 [−4.66, −1.76] (0/8)	+5.26 [+0.39, +11.08] (9/12)	+1.91 [−1.48, +5.85] (9/20)
Flip-delete contrast, $J(P_{\text{high}})$			
Accuracy (pp)	+3.43 [+2.06, +5.00] (8/8)	+5.94 [+3.00, +9.00] (10/12)	+4.93 [+3.04, +6.90] (18/20)
Bal. acc (pp)	+5.50 [+1.05, +12.81] (7/8)	+4.49 [+2.33, +6.72] (10/12)	+4.89 [+2.41, +8.22] (17/20)
Macro-F1 (pp)	+5.86 [+1.69, +12.72] (8/8)	+6.48 [+3.69, +9.64] (12/12)	+6.23 [+3.52, +9.51] (20/20)
True-class prob. ($\times 10^{-2}$)	+12.06 [+8.27, +15.35] (8/8)	+10.42 [+7.24, +13.28] (12/12)	+11.08 [+8.66, +13.38] (20/20)
NLL (nats $\times 10^{-2}$)	+16.88 [+11.43, +21.79] (8/8)	+20.90 [+14.25, +27.53] (12/12)	+19.29 [+14.72, +23.93] (20/20)

Extended to all three cohorts, the secondary-metric picture is both metric- and cohort-dependent rather than a metric-independent property of the high-influence sets. On the original cohort the NLL and true-class-probability excesses are negative with intervals excluding zero ($\Gamma_{\text{flip}} = -3.11$ and -2.76 in the table's units), while accuracy is positive but spans zero; on the expansion cohort the NLL excess reverses to $+5.26$ $[+0.39, +11.08]$ $(9/12)$ and accuracy is supported, while true-class probability stays negative and balanced accuracy and macro-F1 span zero; on the combined cohort accuracy keeps a supported excess, NLL is positive but inconclusive, and true-class probability remains negative with an interval excluding zero. No secondary metric enters the verdicts; the Accuracy row is the prespecified outcome.

G.3 Hub-definition grid sensitivity

On a fixed small grid ($k \in \{3, 5, 10\} \times f_{\text{hub}} \in \{10\%, 15\%, 20\%\}$; all 9 combinations reported at once, Table 25), we recompute on the *original 5-dataset* single-point scope the containment of each hub set in the primary hub set (k=5, 15%), the hub-vs-rest mean single-point damage difference, and the hubness–D_i Spearman (with the oriented importance $H = -h$). The grid uses the split-wise-then-mean aggregation of this sensitivity analysis, so its k=5 values need not exactly equal the pooled main-study estimates of Tables 10 and 11; the pooled estimator remains the primary pointwise estimator. Across all 9/9 combinations: hub-set containment in the primary set 0.64–1.00; hub-vs-rest mean damage differences within $[-0.05, +0.06]$pp; over the reported cells macro Spearman $|\rho| \leq 0.019$, and on D^{NLL} $-0.012 \leq \rho \leq -0.003$ with a constant 2/5 positive direction. The negative evidence for simple hubness strengthens across these 9 definitions; none of the tested settings produces a strong pointwise association.

Table 25: Hub-definition grid (cited in §G): containment of each $h_i(k)$ hub set in the primary set under top-f truncation (containment $= |P_{\text{candidate}} \cap P_{\text{primary}}| \,/\, |P_{\text{candidate}}|$; 1.0 means the candidate set is fully contained in the primary set, i.e. its truncation stays inside the primary top-15% set), set-level hub-vs-rest single-point damage difference (Accuracy, pp) and hubness–D_i Spearman (macro mean; direction count = positive datasets / 5; computed with the oriented importance $H = -h$). Values use the split-wise-then-mean aggregation of this sensitivity analysis, so the k=5 row need not exactly equal the pooled main-study estimates of the primary pointwise estimator (Tables 10 and 11). Primary definition k=5, f=15% (bold row). Tie-corrected D^{Acc} correlations are reported for the primary k=5 definition; for $k \in \{3, 10\}$ the corresponding per-point $h_i(k)$ vectors were not retained, so tie-corrected D^{Acc} correlations are unavailable and are shown as em dashes. The tie convention matters only for D^{Acc}, whose stored values are 99% duplicated; the stored D^{NLL} and D^{prob} vectors contain 0 and 30 duplicate values in 6,051, so the two conventions agree on those columns of every row to within 2×10^{-4}.

k	f_{hub}	containment	hub−rest	$\rho_{D^{\text{NLL}}}$	$\rho_{D^{\text{prob}}}$	$\rho_{D^{\text{Acc}}}$
3	10%	0.91	+0.06pp	−0.003 (2/5)	0.000 (2/5)	—
3	15%	0.82	−0.01pp	−0.003 (2/5)	0.000 (2/5)	—
3	20%	0.71	−0.01pp	−0.003 (2/5)	0.000 (2/5)	—
5	10%	1.00	−0.01pp	−0.007 (2/5)	−0.004 (2/5)	−0.019 (1/5)
5	**15%**	**1.00**	−0.02pp	−0.007 (2/5)	−0.004 (2/5)	−0.019 (1/5)
5	20%	0.75	−0.03pp	−0.007 (2/5)	−0.004 (2/5)	−0.019 (1/5)
10	10%	0.89	−0.05pp	−0.012 (2/5)	−0.018 (1/5)	—
10	15%	0.77	0.00pp	−0.012 (2/5)	−0.018 (1/5)	—
10	20%	0.64	+0.01pp	−0.012 (2/5)	−0.018 (1/5)	—

G.4 Matching sensitivity: exposure definition

The primary W matching targets post-corruption forward exposure (§A.2). As a matching sensitivity we recompute the same matching algorithm with clean forward per-point exposure (greedy, per-class quotas, tie-break and the ≤ 0.10 target unchanged) and rerun the wrong-position intervention: the any-split valid datasets change from balance-scale, iris, wine (n=3) to credit-g (n=1), and credit-g's position advantage becomes -3.40pp under the clean definition (primary definition: $+1.00$pp for that dataset); per-dataset comparison in Table 26. Reading: G_{position} estimates are sensitive to the matching operationalization, and neither definition constitutes positive evidence for a position advantage, consistent with and reinforcing the "position-specific repair unsupported" verdict of Appendix E. The post-corruption-exposure definition remains primary.

Table 26: Matching sensitivity (cited in §G): per-dataset readings after recomputing W matching with clean exposure. G_{position} is the mean over 5 splits (pp); W overlap is the position agreement between clean-W and primary W. Because exposure validity is defined at the split level, the five-split means in this table are descriptive rather than estimates restricted to exposure-valid units; Table 32 reports the corresponding summaries restricted to exposure-valid split units.

Dataset	primary valid	clean valid	G_{pos} primary (pp)	G_{pos} clean (pp)	W overlap
balance-scale	True	False	+0.48	−6.56	0.62
breast_cancer	False	False	+0.53	−2.98	0.52
credit-g	False	True	+1.00	−3.40	0.51
diabetes	False	False	+0.39	−0.65	0.61
iris	True	False	+0.00	−0.67	0.56
tic-tac-toe	False	False	+4.06	−4.27	0.57
vehicle	False	False	+0.71	−3.88	0.78
wine	True	False	+0.56	−0.56	0.52

G.5 Dataset-exclusion robustness

Excluding a dataset and re-running the same estimator is a CPU re-aggregation of the stored per-dataset values, so the exclusion grid is complete rather than searched.

One fixed exclusion. Excluding `steel-plates-fault` — whose own excess is $+0.94$pp, far below the cohort's largest values — leaves the TabPFNv2 combined-cohort excess at $\Gamma_{\text{flip}} = +2.71$pp

$[+0.81, +4.92]$ with $13/19$ datasets direction-consistent. The estimate, the interval and the formal verdict (Support) are unchanged: nothing disappears when that dataset is removed.

Every single deletion. Deleting each of the 20 datasets in turn, the point estimate spans $+2.05$ to $+2.91$pp (full cohort $+2.62$pp) and the CI lower bound spans $+0.48$ to $+1.10$pp, so no deletion moves the interval onto or across zero; the direction count stays at $13/19$ or $14/19$, satisfying the two-thirds criterion; and no deletion changes the verdict (Table 27). The largest single-deletion change is 0.57pp (removing `pc4`), so the average effect is not driven by one dataset.

Table 27: Exclusion robustness of the TabPFNv2 combined-cohort high-vs-matched-random excess Γ_{flip} (accuracy points). Each row re-applies the same estimator (macro mean over the remaining datasets, dataset-level paired bootstrap, seed 20260908) and the three-state criterion (direction fraction $\geq 2/3$); nothing is selected on the outcome, and every single-dataset deletion is listed.

Cohort	Γ_{flip} (pp)
All 20 datasets	**+2.62** $[+0.78, +4.65]$ (14/20)
without `steel-plates-fault`	**+2.71** $[+0.81, +4.92]$ (13/19)
Leave one dataset out (20 deletions)	
balance-scale	$+2.67$ $[+0.74, +4.80]$ (13/19)
breast_cancer	$+2.65$ $[+0.71, +4.79]$ (13/19)
credit-g	$+2.69$ $[+0.76, +4.82]$ (13/19)
diabetes	$+2.72$ $[+0.79, +4.84]$ (13/19)
iris	$+2.77$ $[+0.84, +4.87]$ (14/19)
tic-tac-toe	$+2.73$ $[+0.79, +4.84]$ (13/19)
vehicle	$+2.77$ $[+0.84, +4.87]$ (14/19)
wine	$+2.85$ $[+0.97, +4.92]$ (14/19)
mfeat-morphological	$+2.80$ $[+0.91, +4.89]$ (14/19)
cmc	$+2.70$ $[+0.79, +4.84]$ (13/19)
analcatdata_dmft	$+2.67$ $[+0.77, +4.81]$ (13/19)
pc4	$+2.05$ $[+0.48, +3.87]$ (13/19)
kc2	$+2.16$ $[+0.48, +4.16]$ (13/19)
banknote-authentication	$+2.26$ $[+0.51, +4.33]$ (13/19)
blood-transfusion-service-center	$+2.31$ $[+0.52, +4.41]$ (13/19)
qsar-biodeg	$+2.60$ $[+0.69, +4.81]$ (13/19)
car	$+2.85$ $[+0.99, +5.01]$ (14/19)
steel-plates-fault †	$+2.71$ $[+0.81, +4.92]$ (13/19)
segment	$+2.51$ $[+0.65, +4.70]$ (13/19)
climate-model-simulation-crashes	$+2.91$ $[+1.10, +5.02]$ (14/19)

† the 19-dataset exclusion discussed in the text. Point estimates span $+2.05$ to $+2.91$pp over the 20 deletions (full cohort $+2.62$pp); the largest single-deletion change is 0.57pp (`pc4`); CI lower bounds span $+0.48$ to $+1.10$pp, so no deletion removes the positive interval. Unit: accuracy percentage points; the reported verdict always uses the same rule, not a per-deletion criterion.

G.6 Query-level decomposition of the metric divergence

The set-level cells store the per-query class probabilities of the clean run, the hub-flip run and the ten matched-random flip runs. The three damage measures therefore decompose query by query with no new inference: D_{Acc} as the 0/1 change of the argmax, D_{NLL} as the change in $-\log p_y$, and the true-class probability as p_y itself. Their query means reproduce the set-level excesses exactly — the Accuracy row $+0.49$pp $[-0.33, +1.23]$, NLL -3.11×10^{-2} and true-class probability -2.76×10^{-2} are the values of Table 23 — which lets the divergence be read per query group.

Three groupings fixed in advance (clean correctness, terciles of the pooled clean margin, and the most frequent label as the majority class) are reported in Table 28. No single query-level grouping explains the divergence. The true-class-probability excess is negative in every group of both cohorts, and the NLL excess is negative in every group of the original cohort; on the combined cohort the NLL excess turns positive in five of the eight groups, so the group-level structure is itself cohort-dependent. The Accuracy excess changes sign across groups in both cohorts (positive for clean-correct and middle-margin queries, negative for low-margin ones). In particular the Accuracy excess is *not* concentrated in low-margin queries, so the divergence is not simply a decision-boundary effect, and the minority/majority split does not separate the two families either: the sign difference between the $0/1$ measure and the continuous measures survives inside both class-frequency groups.

Table 28: Query-level decomposition of the set-level excess, from the saved per-query probabilities, on both the original cohort and the combined 20-dataset cohort. Each cell is the query mean (then dataset macro) of $\Gamma_{\text{metric}}(q) = D^{\text{hub}}_{\text{metric}}(q) - \overline{D^{\text{rand}}_{\text{metric}}}(q)$ restricted to a fixed query group; units are pp for Accuracy and $\times 10^{-2}$ for NLL (nats) and the true-class probability. Groups are fixed in advance (clean correctness; terciles of the pooled clean margin; the most frequent label as the majority class), not chosen by outcome.

Query group	Γ_{Acc} (pp)	Γ_{NLL} (nats $\times 10^{-2}$)	Γ_{prob} ($\times 10^{-2}$)
original cohort, 8 datasets			
all queries (reference)	$+0.49$ (5/8)	-3.11 (0/8)	-2.76 (0/8)
clean-correct queries	$+0.88$ (5/8)	-3.42 (1/8)	-2.95 (0/8)
clean-incorrect queries	-2.32 (3/7)	-1.14 (3/7)	-1.53 (2/7)
low clean margin (bottom tercile)	-2.61 (4/8)	-1.35 (3/8)	-0.62 (3/8)
middle clean margin	$+1.23$ (5/8)	-2.63 (2/8)	-2.43 (2/8)
high clean margin (top tercile)	$+0.74$ (3/7)	-4.70 (2/7)	-4.45 (1/7)
majority-class queries	$+2.23$ (4/8)	-1.87 (2/8)	-2.19 (2/8)
minority-class queries	-3.40 (1/8)	-4.43 (0/8)	-3.30 (0/8)
combined cohort, 20 datasets			
all queries (reference)	$+2.62$ (14/20)	$+1.91$ (9/20)	-1.57 (6/20)
clean-correct queries	$+3.73$ (14/20)	$+2.39$ (10/20)	-1.57 (6/20)
clean-incorrect queries	-2.27 (8/18)	$+1.01$ (9/18)	-0.68 (7/18)
low clean margin (bottom tercile)	-2.96 (8/19)	-0.78 (8/19)	-0.80 (8/19)
middle clean margin	$+0.25$ (9/19)	-4.90 (6/19)	-4.98 (5/19)
high clean margin (top tercile)	$+4.93$ (8/16)	$+6.55$ (7/16)	-0.62 (4/16)
majority-class queries	$+3.53$ (11/20)	$+1.99$ (8/20)	-1.53 (5/20)
minority-class queries	-3.07 (8/20)	-1.12 (8/20)	-2.16 (7/20)

G.7 Pointwise descriptive robustness: binary subset and Accuracy ties

Binary subset. For a multiclass context point the single-point intervention evaluates one deterministic wrong-target realization, so the measured damage can carry a target-realization component. Restricting the pointwise analysis to the five datasets with two classes removes that component entirely. Table 29 reports the same estimator on that subset: the reliable-ranker criterion remains unmet for both readouts. TabPFNv2's oriented hub importance reaches $\rho_{\text{NLL}} = -0.117\ [-0.179, -0.054]$ with $0/5$ datasets direction-consistent, and TabICLv2's prototype drift reaches $+0.088\ [+0.004, +0.204]$ with $4/5$ datasets direction-consistent; its point estimate stays well below the $\rho \geq 0.20$ bar, so here the criterion is unmet rather than association absent. Weak pointwise correspondence therefore also persists on the binary subset, where the wrong target is deterministic.

Accuracy ties. Table 30 records how often a single-point flip leaves the query-set accuracy unchanged, i.e. how often $D_{\text{Acc}} = 0$. Because D_{Acc} is the difference between two query-set accuracy proportions it is discrete, and exactly zero in most of the study — 57% of all records for TabPFNv2 and 62% for TabICLv2, with some dataset–model cells exceeding 90% — which is why D^{NLL} is the primary pointwise damage measure and Accuracy is retained only as the scale-matched secondary check. A zero value does not require every query prediction to remain unchanged: a flip can move some queries from correct to incorrect and others the other way, with no net change in the query-set accuracy. This is a property of the measure, not a new result.

Table 29: Pointwise correspondence on the binary datasets only (5 of the 9: breast_cancer, credit-g, blood-transfusion-service-center, kc2, climate-model-simulation-crashes). With two classes the wrong target is unique, so the measured damage carries no target-realization component; this is a descriptive robustness read with the same estimator, with no criterion re-designed for the subset and no other subset searched. Criterion met requires $\rho \geq 0.20$, CI lower bound > 0 and at least two thirds of the datasets direction-consistent.

Signal	Damage	ρ	95% CI	dir.
TabPFNv2 H (hubness)	D^{NLL}	-0.117	$[-0.179, -0.054]$	0/5
TabPFNv2 H (hubness)	D^{Acc}	$+0.001$	$[-0.053, +0.073]$	2/5
TabPFNv2 H (hubness)	D^{prob}	$+0.012$	$[-0.102, +0.126]$	3/5
TabPFNv2 M (max attention)	D^{NLL}	$+0.065$	$[-0.029, +0.178]$	3/5
TabPFNv2 M (max attention)	D^{Acc}	$+0.026$	$[-0.051, +0.117]$	3/5
TabPFNv2 M (max attention)	D^{prob}	-0.215	$[-0.266, -0.176]$	0/5
TabPFNv2 $\bar{E}$ (mean exposure)	D^{NLL}	$+0.040$	$[-0.058, +0.179]$	3/5
TabPFNv2 $\bar{E}$ (mean exposure)	D^{Acc}	$+0.026$	$[-0.005, +0.057]$	4/5
TabPFNv2 $\bar{E}$ (mean exposure)	D^{prob}	-0.234	$[-0.407, -0.028]$	1/5
TabICLv2 I^{proto} (prototype drift)	D^{NLL}	$+0.088$	$[+0.004, +0.204]$	4/5
TabICLv2 I^{proto} (prototype drift)	D^{Acc}	-0.025	$[-0.091, +0.032]$	1/5
TabICLv2 I^{proto} (prototype drift)	D^{prob}	-0.028	$[-0.070, +0.015]$	2/5

The reliable-ranker criterion remains unmet on the binary subset for both readouts. TabICLv2's prototype drift shows a small positive association on D^{NLL} whose point estimate stays well below the $\rho \geq 0.20$ bar, so an unmet criterion does not imply the absence of association.

Table 30: Tie structure of the single-point Accuracy damage D_{Acc}: share of single-point flips for which the query-set accuracy is unchanged, i.e. $D_{\mathrm{Acc}} = 0$, over the 3-split pointwise study. Because D_{Acc} is the difference between two query-set accuracy proportions it is discrete, and exactly zero whenever the flip leaves the number of correctly classified queries unchanged; a zero value does not require every query prediction to stay the same, since a flip can move some queries from correct to incorrect and others the other way. This is why D^{NLL} is the primary pointwise damage measure and Accuracy is kept as the scale-matched secondary check; no analysis is redesigned here.

Dataset	TabPFNv2	TabICLv2	records
balance-scale	65.3%	89.2%	1500
breast_cancer	57.8%	56.5%	1365
credit-g	43.9%	50.7%	2400
iris	81.9%	96.1%	360
wine	94.1%	70.2%	426
blood-transfusion-service-center	77.6%	87.0%	1794
kc2	64.8%	74.3%	1251
analcatdata_dmft	26.4%	38.1%	1911
climate-model-simulation-crashes	59.7%	36.0%	1296
Pooled	56.9%	62.2%	12303 / 12303

H Comparison with Biloš et al. (2026)

Table 31: Study-design comparison with Biloš et al. (2026).

Item	Biloš et al. (2026)	Ours
Question	mechanism-grounded attack effectiveness (which readout-targeted perturbation hurts; isolate readouts from refit baselines)	position-specific causal attribution of wrong supervision (does position matter; do the tested readout-aligned importance scores rank single-point damage)
Attack	hub poison (flip labels of context points with smallest average kNN distance to all others), rank warp, noise padding, centroid/boundary injection, SVD hide, null-space PGD; 9 attacks in 4 families	one corruption type: wrong labels at a fixed 15% budget (flip vs. deletion; high-influence vs. matched-random; single-point flips)
High-influence selection	hubs: top 15%, "rep-cosine kNN degree at L9" (k not stated in published text)	TabPFNv2: L9 cosine kNN centrality, k=5, top 15%; TabICLv2: leave-one-out prototype drift, top 15%
Wrong-label destination	"labels flipped" (destination not stated)	uniform other class, applied to the high set and to every class-stratified matched-random set under per-set deterministic RNG streams
Random matched control	none (contrasts vs. clean and vs. refit Ridge/XGBoost/MLP)	matched-random sets, R=10, disjoint, class-stratified
Flip–delete decomposition	no	yes ($J = D_{\text{flip}} - D_{\text{delete}}$)
Single-point damage	no (set-level only)	yes: per-point causal matrix over the single-point cohort (9 datasets × 3 splits; 12,303 flips per model arm)
Dataset cohort	49 classification + 10 regression; attack headline on common 24-classification grid × 5 seeds (their Tab. 3); broader sweep 49 × 10 seeds	20 classification datasets × 5 stratified splits (8 original + 12 expansion); single-point scope 9 × 3 splits
Splits/seeds	80/20 stratified per (dataset, seed); seeds $0, \ldots, N-1$; partitions reused across models and baselines	80/20 stratified per split seed $\in \{0, \ldots, 4\}$; partitions reused across both model arms
Preprocessing / model	official released checkpoints (TabPFNv2 12-block; TabICLv2), fp32, deterministic cuDNN, TF32 off	official released checkpoints and official pipelines, fp32 deterministic, TF32 off
Reported estimand	mean accuracy drop (post-attack − clean) with 95% paired-bootstrap CI over datasets	macro accuracy loss in pp with dataset-level paired-bootstrap 95% CI and direction counts; three-state criterion (Support/Reject/Inconclusive)

Table 31 compares the attack and evaluation design of the direct predecessor (Biloš et al., 2026) (extracted from its published appendix text) with the present study. Where their protocol is silent in the published text, the cell says "not stated"; no cell is filled by assumption.

H.1 Repair case study: exposure-valid split units

Table 32: Repair case study re-aggregated on exposure-valid split units (the splits whose exposure ratio in Table 17 is ≤ 0.10). Each entry is the mean over that dataset's valid splits; accuracy columns are accuracies and R_{correct}, R_{wrong}, G_{position} are accuracy-unit differences, as in Table 16. The unit is the split, not the dataset: five of the eight datasets contribute no valid split and therefore do not appear.

Dataset	valid splits	Poison	Correct	Wrong	Clean Same	R_{correct}	R_{wrong}	G_{position}
balance-scale	5/5	0.901	0.904	0.899	0.970	+0.003	-0.002	+0.005
iris	4/5	0.967	0.958	0.967	0.967	-0.008	+0.000	-0.008
wine	4/5	0.986	0.993	0.986	1.000	+0.007	+0.000	+0.007