CART: CONTEXTUAL CELL RELIABILITY FOR ROBUST TABULAR PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world tabular data is rarely uniformly reliable: individual entries (cells) can be corrupted or mis-recorded while the rest of a record remains trustworthy, yet most robust tabular learning methods act at the sample level (train the whole model to be robust), the feature level (weight entire columns), or attempt to reconstruct every cell indiscriminately. We argue that none of these directly answers the question a predictor implicitly has to resolve at inference time: *how much should this particular observed cell be trusted for this particular record?* We propose CART, an architecture that (i) estimates a contextual, per-cell reliability score from full-record attention, (ii) reconstructs a true leave-one-out (LOO) counterfactual representation of every cell using an attention query that never observes the cell itself, and (iii) adaptively fuses the observed and reconstructed representations according to the learned reliability before classification. Across eight tabular benchmarks, three seeds, and six corruption severities, CART attains the best noise-robustness AUC on 7 of 8 datasets and the best or tied-best accuracy at 20% corruption on 6 of 8 datasets, while remaining competitive on clean data. A seven-way mechanistic ablation shows that the improvement is not explained by indiscriminate reconstruction, a fixed repair ratio, a random gate, or even a gate with the correct value distribution but permuted cell assignment (two-sided Wilcoxon signed-rank test across 8 datasets, $p = 0.0078$ in all three comparisons): the gain requires assigning the *correct* reliability score to the *correct* cell. Compared with reconstruction-oriented baselines built on the same tokenizer (a denoising autoencoder, a transformer autoencoder without LOO, and a variational autoencoder), CART attains the best cell-level reliability detection on roughly half of the datasets and is a close competitor elsewhere, showing that generic reconstruction quality and cell-reliability detection are related but distinct capabilities. Finally, CART's robustness advantage persists, though it does not always dominate, under corruption magnitudes and feature shifts not seen during training.

1 INTRODUCTION

Tabular data used in production machine learning systems is rarely uniformly reliable. Sensor glitches, data-entry mistakes, and accidental or adversarial corruption typically affect individual *cells* rather than entire records or columns: a patient's income field may be mistyped while every other field, and every other patient's income field, remains correct. This creates a fine-grained reliability problem that existing robust tabular learning methods do not directly address.

Existing approaches fall into three families, none of which targets this problem directly. *Robust prediction* methods (augmentation, adversarial training, QuAIL (Mattia Sabella & Pietro Pinoli, 2026)) train the whole predictor to tolerate corruption but do not expose which specific cells were untrustworthy. *Reconstruction and imputation* methods attempt to reconstruct every cell, whether or not it needs repairing, conflating "cells that can be reconstructed" with "cells that should be replaced." *Feature- or sample-level reliability* methods assign one score to an entire column (QuAIL) or record (sample-level anomaly detection), which is too coarse: a column can be reliable for most records and unreliable for one. None of these directly answers:

> *How much should this particular observed cell be trusted for predicting the label of this particular record?*

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.

We call this *cell reliability*, distinct from two related notions. *Feature importance* (SHAP (Lundberg & Lee, 2017), Local MDI+ (Liang et al., 2026)) measures a feature's contribution to a prediction, not whether its value should be trusted. *Cell-level anomaly detection* (Wang et al., 2026) measures statistical unusualness, which is neither necessary nor sufficient for untrustworthiness: a genuinely large age value can be unusual yet trustworthy, while a data-entry error can lie well within a column's normal range and still be wrong.

CART. We propose CART (Contextual cell-level Adaptive Repair for Tabular data), which operationalizes cell reliability with a simple decomposition: a *trust branch* estimates reliability from full-record context; a *repair branch* reconstructs a counterfactual cell representation using only the other cells, via an attention query architecturally prevented from observing the cell's own value (true leave-one-out, LOO); and *adaptive fusion* combines the observed and repaired representations in proportion to the learned reliability before classification. The repair pathway makes reconstruction meaningful (it cannot copy the observed value); the trust pathway decides, per cell and record, how much to use it.

A natural concern is circularity: if reliability decides how much to trust a cell, and the model is also trained to be robust, is it doing anything beyond curve-fitting the training corruption? We address this by supervising the trust head, in an early stage, directly with the synthetic corruption mask and the repair branch with an explicit reconstruction loss (both external to the student's own predictions), then adding a later-stage consistency loss against a frozen EMA teacher, so trust is guided by an external teacher target rather than by the student's own prediction confidence; "label-free" always means *without human-annotated real-world corruption*, and no mask is available at test time.

Rather than assume this decomposition works, we test it directly. CART improves robust accuracy over five prediction-oriented baselines across eight datasets and three seeds; a seven-way mechanistic ablation shows raw inputs, indiscriminate repair, a fixed or random gate, and a gate with the right values but permuted across cells all fail to reproduce CART's gain, confirmed by a dataset-level statistical test. We also isolate the trust-and-repair decomposition from mere encoder capacity via reconstruction-oriented baselines sharing CART's tokenizer, and evaluate generalization to corruption not seen during training. Throughout, "CART" denotes the default sampled-rate model, except CART-fixed, a training-schedule ablation (Section 5.6).

Our contributions: (1) we formulate **task-conditioned cell reliability** as distinct from feature importance, sample-level anomaly detection, and generic reconstruction, with a concrete architectural operationalization; (2) we introduce **CART**, combining a contextual cell-trust branch, a true leave-one-out repair branch, and adaptive reliability-guided fusion, trained with a two-stage teacher–student scheme that uses external corruption-mask supervision and a frozen-teacher consistency target to guide trust learning; (3) across 8 datasets and 3 seeds, CART consistently improves robustness under cell-level corruption relative to five strong baselines, while remaining competitive (though not always best) on clean data; (4) a seven-way mechanistic ablation, a validation-selected fixed-gate comparison, and a dataset-level statistical test show the gain is attributable specifically to *adaptive, correctly-assigned, cell-level* reliability gating, not reconstruction alone, a favorable fixed hyperparameter, or a randomly distributed gate; (5) CART's reliability signal differs meaningfully from generic reconstruction quality (vs. a DAE, a transformer autoencoder without LOO, and a VAE sharing its tokenizer), and its robustness advantage largely persists under corruption magnitudes and feature shifts not used during training.

2 Related Work

2.1 Robust Tabular Learning

Gradient-boosted trees (CatBoost (Dorogush et al., 2018), XGBoost (Chen & Guestrin, 2016)) remain strong predictors on clean tabular data and can still outperform deep learning architectures (Grinsztajn et al., 2022); deep tabular architectures (MLP (Rumelhart et al., 1986), FT-Transformer (Gorishniy et al., 2021), TabNet (Arik & Pfister, 2021), SAINT (Somepalli et al., 2021)) are typically trained without an explicit mechanism for identifying unreliable inputs, and robustness is usually pursued through data augmentation alone. QuAIL (Mattia Sabella & Pietro Pinoli, 2026) is the closest prior work: it learns a feature-level reliability prior that modulates embeddings before classification. The key difference is granularity – QuAIL's reliability is a property of a *column*,

shared across all records, whereas CART estimates reliability per *cell*, conditioned on the rest of that record, and repairs (rather than only down-weights) the cell via leave-one-out reconstruction.

2.2 Cell Corruption and Anomaly Detection

A parallel line of work moves anomaly detection from the sample level to the cell level (Wang et al., 2026). This is a valuable but different signal from reliability (Section 6): a cell can be simultaneously anomalous-and-trustworthy or unremarkable-and-corrupted, because anomaly reflects the data distribution, not whether a value should be trusted for a specific task.

2.3 Reconstruction and Imputation

Denoising autoencoders (Vincent et al., 2008), GAIN (Yoon et al., 2018), VIME (Yoon et al., 2020), and statistical imputers such as MissForest (Stekhoven & Bühlmann, 2012) and HyperImpute (Jarrett et al., 2022) learn tabular representations via reconstruction or imputation, with contrastive self-supervision (SCARF (Bahri et al., 2022)) as a related direction; reconstruction error is sometimes used as an implicit corruption signal. GAIN and VIME are optimized for missing-value imputation rather than corruption detection under a fixed classification objective and use different encoder architectures, complicating direct comparison. We instead compare against baselines that reuse CART's own tokenizer and Transformer encoder – a DAE classifier, a transformer autoencoder without leave-one-out repair (TabAE-E2E), and a VAE (Kingma & Welling, 2014) – so any advantage is attributable to the trust-and-repair mechanism, not encoder capacity.

2.4 Reliability, Importance, and Distribution Shift

Instance-specific importance methods (SHAP (Lundberg & Lee, 2017), Local MDI+ (Liang et al., 2026)) quantify a feature's contribution to a prediction, which is empirically distinct from reliability (Section 6). TabFSBench (Cheng et al., 2025) formalizes feature-shift robustness as an external corruption protocol, and tabular test-time adaptation (He & Shi, 2026) lets the model adapt using target-domain data; CART instead targets robust prediction *without* target-domain adaptation. Selective classification and conformal prediction under shift (Liang et al., 2024; Lee, 2026) address a complementary question – when to abstain – which we treat as a downstream use of a reliability score, not a substitute for estimating it.

3 The CART Architecture

Figure 1 gives an overview of CART; the remainder of this section describes each component.

3.1 Problem Formulation

Let $x_i = [x_{i1}, \ldots, x_{id}]$ denote a d-dimensional tabular record (numerical and categorical features), and let $\tilde{x}_i$ denote a possibly corrupted observation of x_i, in which an unknown subset of cells has been altered by noise, replacement, or other corruption mechanisms. Given only $\tilde{x}_i$, CART must produce a class prediction $\hat{y}_i = f(\tilde{x}_i)$ that is robust to the (unknown, unlabeled at test time) corruption, while also producing a per-cell reliability estimate $r_i = g(\tilde{x}_i) \in [0, 1]^d$, where r_{ij} estimates how much the observed value $\tilde{x}_{ij}$ should be trusted when predicting y_i. At training time, synthetic corruption masks are used as (a) an auxiliary detection target for g and (b) reconstruction targets for the repair branch; they are never available to the model at test time, when r_{ij} is inferred purely from context.

3.2 Cell Tokenization and Contextual Trust

Each record is represented as a sequence of d feature tokens $z = [z_1, \ldots, z_d]$ from a shared tokenizer (FT-Transformer-style): a per-feature linear projection for numerical features, an embedding lookup for categorical features, each summed with a feature-identity and a feature-type embedding. A stack of two standard Transformer blocks (Vaswani et al., 2017) (full self-attention over all d tokens) produces a context-aware representation c^{trust}, and a small MLP head maps the token-context pair to

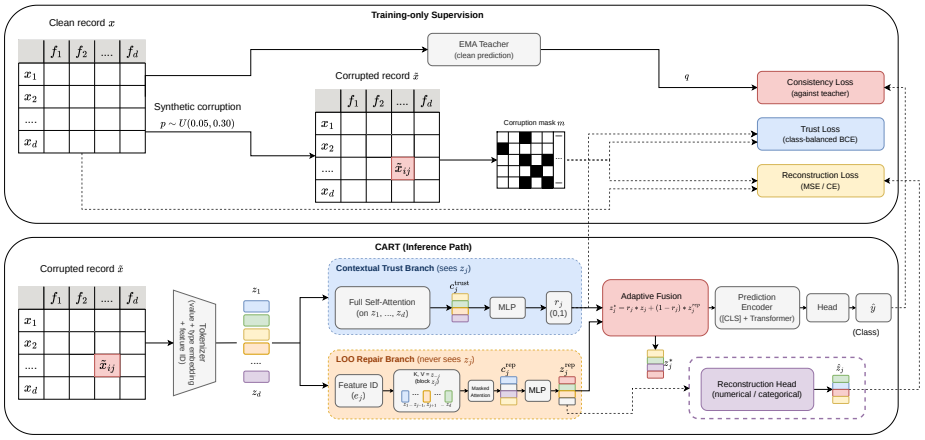

Figure 1: CART overview. A trust branch estimates a per-cell reliability score r_j from full self-attention over z. The repair branch is architecturally constrained to leave-one-out: its query is the feature-identity embedding e_j (not z_j), its keys/values are $z_{\setminus j}$, and cell j is additionally masked out of its own attention, so z_j^{rep} cannot depend on the observed value of cell j. The two are adaptively fused before the prediction encoder; reconstruction heads on the repair branch provide the cell-recovery diagnostics of Section 5.2.

a trust logit,

$$r_j = \sigma\big(\text{MLP}_{\text{trust}}([z_j; c_j^{\text{trust}}])\big). \tag{1}$$

Because the trust branch can attend to the (possibly corrupted) token z_j itself, it can learn cues such as "this value is inconsistent with the rest of the record."

3.3 True Leave-One-Out Repair

The repair branch produces a reconstruction of cell j that is architecturally prevented from using the observed value z_j, reminiscent of masked-token prediction in language models (Devlin et al., 2019) adapted to per-feature tabular attention. Concretely, its query is not z_j but the feature-identity embedding e_j alone; the key/value sequence is z, attended under a mask that additionally blocks j from attending to itself:

$$c_j^{\text{repair}} = \text{Attn}\big(q{=}e_j,\ K{=}V{=}z_{\setminus j}\big), \tag{2}$$

$$z_j^{\text{rep}} = \text{MLP}_{\text{repair}}\big([c_j^{\text{repair}}; e_j]\big). \tag{3}$$

Because the query itself carries no information about the observed cell, z_j^{rep} is a genuine leave-one-out counterfactual: it is the representation the model would construct for cell j if it could only see the rest of the record. Explicit reconstruction heads decode z_j^{rep} back to the original feature space (a scalar for numerical features, a categorical distribution for categorical features); this both regularizes z^{rep} to carry recoverable content and provides the cell-recovery metrics of Section 5.2.

3.4 Adaptive Reliability-Guided Fusion

The trust score r_j is used to fuse the observed and repaired representations of every cell:

$$z_j^\star = r_j \cdot z_j + (1 - r_j) \cdot z_j^{\text{rep}}. \tag{4}$$

A cell deemed reliable ($r_j \to 1$) passes through unchanged; an unreliable cell ($r_j \to 0$) is effectively replaced by its leave-one-out counterfactual. A final two-block Transformer encoder (with a prepended `[CLS]` token) processes $z^\star = [z_1^\star, \ldots, z_d^\star]$, and a classification head produces $\hat{y}$ from the `[CLS]` representation. Reliability and reconstruction are thus not auxiliary tasks competing with prediction; they shape the representation that is fed to the classifier, and their quality is evaluated in Section 5.2 as a diagnostic of whether that representation is meaningful.

3.5 Teacher–Student Training and Corruption Schedule

Defining reliability purely as "whatever makes the model's own predictions stable" risks circularity. CART instead trains in two stages: in stage 1 (30 of 80 epochs), the trust head is supervised directly by the synthetic corruption mask (class-balanced BCE), and the repair branch by the reconstruction losses – neither depends on the student's own predictions. In stage 2, a frozen EMA teacher computes clean-data prediction confidence, which regularizes the student's fused prediction on corrupted inputs via a confidence-weighted consistency loss; the teacher never receives gradients from this loss. This separates *target generation* (teacher confidence, synthetic mask) from *trust learning* (the student). By default CART samples a corruption rate $\rho \sim \mathrm{Uniform}(0.05, 0.30)$ per batch; a training-schedule ablation, CART-fixed, instead uses a fixed rate $\rho = 0.20$ (evaluated in Section 5.6). Unless stated otherwise, "CART" denotes the default sampled-rate model.

3.6 Training Objective

The overall loss is $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_1 \mathcal{L}_{\text{trust}} + \lambda_2 \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{cons}}$ (with $\lambda_3 = 0$ during stage 1), where $\mathcal{L}_{\text{pred}} = \mathrm{CE}(y, \hat{y})$, $\mathcal{L}_{\text{trust}}$ is the detection loss above, $\mathcal{L}_{\text{rec}}$ is a masked MSE (numerical) / cross-entropy (categorical) reconstruction loss on corrupted cells only, and $\mathcal{L}_{\text{cons}}$ is the stage-2 consistency term.

4 Experimental Setup

Datasets. We use 8 tabular binary classification datasets spanning different sizes, feature compositions, and class balances: Adult (Kohavi, 1996), Bank (Moro et al., 2014), Churn (Kaggle, 2020), Credit (Yeh & Lien, 2009), Diabetes (Smith et al., 1988), HELOC (FICO, 2017), Higgs (Baldi et al., 2014), and Magic (Bock et al., 2004), primarily sourced from OpenML (Vanschoren et al., 2013). Each is run with 3 seeds (42, 123, 2024). Details on the task and feature composition of each dataset are provided in Appendix A.

Corruption protocols. CART samples $\rho \sim \mathrm{Uniform}(0.05, 0.30)$ per training batch; CART-fixed uses $\rho = 0.20$ (Section 5.6). Corruption perturbs each selected cell independently: numerical cells get additive Gaussian noise, categorical cells are replaced by a different category drawn uniformly at random. Standard test corruption uses severities $\rho \in \{0, 0.1, \ldots, 0.5\}$, matching the training mechanism. To probe generalization beyond this distribution, we also evaluate (i) *unseen-magnitude* corruption – 20% of numerical cells perturbed by $\mathrm{Uniform}(3, 5)$ additive noise (standardized units, random sign), substantially larger than the $\mathcal{N}(0, 2)$ training noise – and (ii) *feature-shift* corruption, which shifts one entire randomly-selected numerical feature by $\mathrm{Uniform}(2, 4)$, a qualitatively different, non-cell-sparse pattern.

Baselines. *Group A (prediction-oriented):* CatBoost, XGBoost, FT-Transformer, MLP, and QuAIL, trained with the same fixed 20%-corruption augmentation as CART-fixed. *Group B (reconstruction-oriented):* a CatBoost classifier on denoising-autoencoder (DAE) features; TabAE-E2E, which reuses CART's own tokenizer and Transformer blocks with a reconstruction/detection head but *without* leave-one-out repair (its context branch can see the cell it reconstructs); and a VAE classifier. TabAE-E2E in particular isolates the effect of true leave-one-out repair from simply having a Transformer encoder.

Evaluation metrics. Robust prediction: clean accuracy, Acc@20, and NR-AUC (area under the accuracy-vs-severity curve, $\rho \in \{0, \ldots, 0.5\}$, normalized). Cell reliability/recovery: Cell-AUROC and Cell-AUPRC of the learned score against the ground-truth mask, numerical reconstruction MAE, and categorical reconstruction accuracy, computed at $\rho = 0.2$.

Implementation details. All neural models share the optimizer (AdamW, lr 3×10^{-4}, weight decay 10^{-4}) and embedding dimension (32). CART and the neural Group-A baselines (MLP, FT-Transformer, QuAIL) are all trained for exactly 80 epochs to ensure a strictly fair optimization budget. CART uses its two-stage trust/reconstruction/consistency objective, whereas the baselines use their respective standard supervised objectives with corruption augmentation.

Table 1: Overall robust prediction performance under standard cell corruption (mean over 3 seeds, %). "Best other" is the strongest of the five Group-A baselines (CatBoost, XGBoost, FT-Transformer, MLP, QuAIL) on that column; **bold** marks the best value; ties are bolded for both. Full per-model breakdown with std in Appendix B, Table 4.

	Clean Acc.		Acc@20%		NR-AUC	
Dataset	**Best other**	**CART**	**Best other**	**CART**	**Best other**	**CART**
Adult	**83.89**	83.76	81.64	**82.03**	80.75	**80.98**
Bank	**89.88**	89.68	**89.01**	88.94	88.47	**88.66**
Churn	**93.11**	92.76	89.51	**90.40**	88.45	**89.90**
Credit	**81.78**	81.74	81.05	**81.09**	**80.80**	80.71
Diabetes	75.57	**77.30**	**74.14**	**74.14**	71.78	**71.84**
HELOC	**72.87**	72.44	**71.47**	71.09	70.10	**70.39**
Higgs	71.12	**71.34**	67.25	**67.97**	66.30	**66.90**
Magic	86.13	**86.46**	81.83	**82.86**	80.08	**80.60**

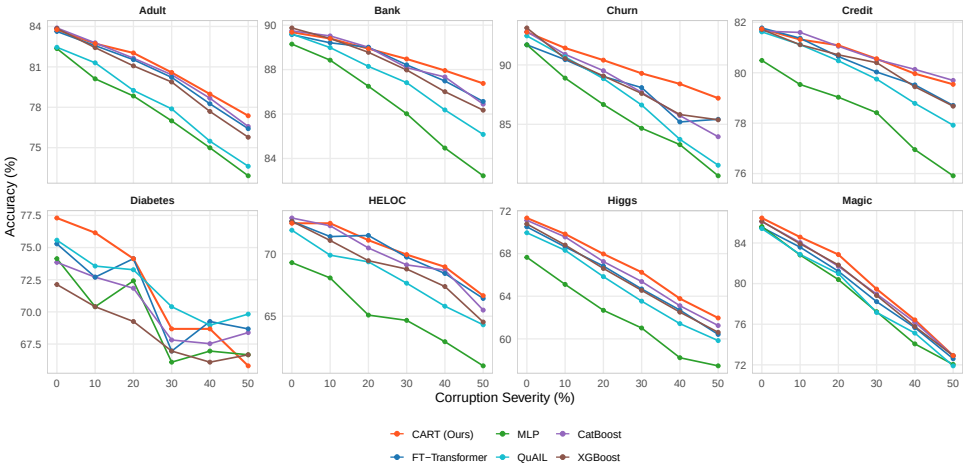

Figure 2: Accuracy vs. corruption severity, averaged over 3 seeds, for all 8 datasets (note the free y-axis per panel). CART (orange) degrades more slowly than the five Group-A baselines (grey) on most datasets as corruption severity increases; this is the trajectory that NR-AUC in Table 1 summarizes into a single number.

5 Results

5.1 Overall Robust Prediction

Table 1 reports clean accuracy, Acc@20, and NR-AUC for CART against the five Group-A baselines. CART attains the best NR-AUC on **7 of 8 datasets**, losing only to CatBoost on Credit (80.71 vs. 80.80, a 0.09-point gap), and the best or tied-best Acc@20 on **6 of 8 datasets** (tying MLP exactly on Diabetes at 74.14). CART is the best clean-accuracy model on 3/8 datasets (Diabetes, Higgs, Magic) and is within 1–2 points of the best elsewhere; it is not intended, and does not need, to be a universally better clean-data classifier – its advantage should manifest specifically as corruption increases. Figure 2 makes this pattern directly visible: CART's accuracy-vs-severity curve is flatter than the baselines' on most datasets, which is exactly what NR-AUC is designed to summarize into a single number.

5.2 Cell Reliability and Reconstruction

Table 1 shows CART improves robust prediction, but does not show *why*. We first ask whether CART's internal reliability signal is meaningful. Figure 3 plots the mean learned trust score for clean vs. corrupted cells on each of the 8 datasets (computed on held-out corrupted data from the trained CART checkpoints). Clean-cell trust exceeds corrupted-cell trust on **all 8 datasets** (e.g., 0.86

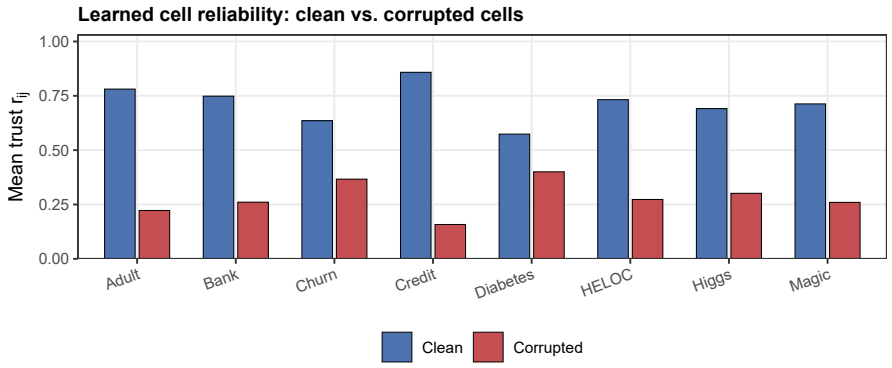

Figure 3: Mean learned trust score for clean vs. corrupted cells, per dataset. Clean cells receive higher trust than corrupted cells on every dataset, despite r_{ij} being inferred without test-time corruption labels.

Table 2: Cell-level reliability detection and recovery, averaged over 8 datasets (Cat. Acc. averaged over the 3 datasets with categorical features). Cell-AUROC/AUPRC (↑), numerical reconstruction MAE (↓), categorical reconstruction accuracy (%, ↑); best value in **bold**, second-best underlined. Full per-dataset breakdown with std in Appendix B, Table 5.

Model	Cell-AUROC	Cell-AUPRC	Num. MAE	Cat. Acc.
CART	**0.863**	**0.951**	**0.557**	**69.74**
DAE	0.840	0.936	0.637	45.77
TabAE	0.858	0.947	0.678	47.75
VAE	0.812	0.927	0.581	50.49

vs. 0.16 on Credit, 0.78 vs. 0.22 on Adult; the narrowest separation is on Diabetes, 0.57 vs. 0.40), even though r_{ij} receives no corruption label at test time.

5.3 Comparison with Reconstruction-Oriented Baselines

A natural objection is that a generic reconstruction objective might already give an equally good reliability signal, e.g., by thresholding reconstruction error. Table 2 compares CART against DAE, TabAE-E2E, and VAE. CART attains the best Cell-AUROC on 4/8 datasets, Cell-AUPRC on 6/8, lowest numerical MAE on 5/8, and best categorical accuracy on all 3 categorical datasets. TabAE-E2E – sharing CART's tokenizer/Transformer but reconstructing *without* leave-one-out masking – is the strongest competitor, occasionally exceeding CART on Cell-AUROC (Higgs, and a near-tie on Magic); DAE is occasionally competitive on Diabetes/Churn. We do not claim CART dominates every metric on every dataset: the defensible claim is that CART is consistently competitive with, and frequently better than, reconstruction baselines on the same tokenizer, while also producing an explicit, actionable reliability score that a pure autoencoder does not (Appendix B, Figure 5 and Table 5).

5.4 Why Does Adaptive Cell Reliability Work?

Sections 5.1–5.3 establish *that* CART is robust and produces a meaningful reliability signal. This section asks the harder question: is the gain caused by *adaptive, per-cell* gating, or merely by having a repair mechanism or some arbitrary gate?

Seven-way ablation. Using the trained CART checkpoints for each dataset and seed, we construct $z^\star$ seven ways at inference time on the 20%-corruption test split: (1) *Raw* ($z^\star = z$); (2) *Repair-only* ($z^\star = z^{\text{rep}}$); (3) *Fixed gate*, $r_j \equiv 0.8$ (an inference-time constant, distinct from the CART-fixed training-schedule ablation of Section 5.6); (4) *Random gate*, $r_j \sim \text{Uniform}(0, 1)$ per cell; (5) *Permuted trust* – the model's own learned trust *values*, shuffled across cells; (6) *CART* – the model's

Table 3: Seven-way mechanistic ablation: accuracy (%) at 20% corruption, evaluated using the trained CART checkpoints for each dataset and seed, with results averaged over 3 seeds. Best value per row (excluding the oracle-mask intervention) in **bold**.

Dataset	Raw	Repair	Fixed	Random	Perm.	CART	Oracle
Adult	79.5	78.4	80.2	80.1	80.2	**82.0**	82.7
Bank	88.0	87.5	88.1	88.1	88.2	**88.9**	88.0
Churn	86.9	58.6	86.9	85.8	86.4	**90.4**	86.9
Credit	78.3	77.7	79.2	79.4	78.7	**81.1**	80.7
Diabetes	71.0	44.8	72.1	64.9	69.8	**74.1**	72.7
HELOC	64.0	50.3	67.8	61.9	66.4	**71.1**	65.7
Higgs	57.8	52.8	60.4	56.3	57.5	**68.0**	61.1
Magic	65.4	35.2	69.8	59.8	78.9	**82.9**	65.5
Mean	*73.9*	*60.7*	*75.6*	*72.0*	*75.8*	***79.8***	*75.4*

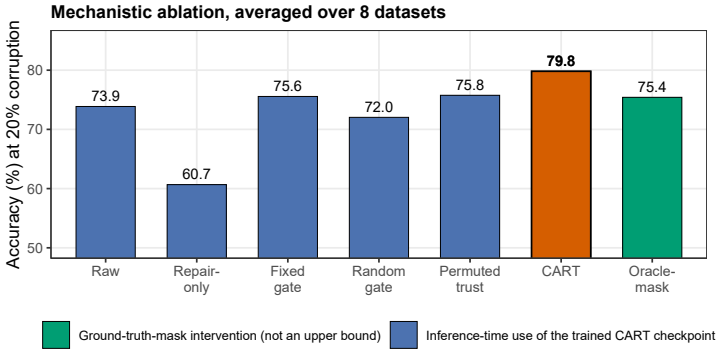

Figure 4: Mechanistic ablation, averaged over 8 datasets. CART's learned, correctly-assigned per-cell gate (orange) outperforms every alternative use of the same trained checkpoint (blue), including a gate with the identical value distribution but permuted cell assignment; the oracle-mask intervention (green) is a ground-truth-mask substitution, not an upper bound.

own correctly-assigned r_j; (7) *Oracle-mask* – r_j replaced by the ground-truth mask (not an upper bound: it discards the learned soft fusion for a hard, possibly miscalibrated substitution). Table 3 and Figure 4 report per-dataset and mean accuracy.

Three findings stand out. First, *Repair-only* is often catastrophic (Churn 58.6, Diabetes 44.8, Magic 35.2): indiscriminately replacing every cell with its counterfactual is actively harmful – reconstruction helps only when applied selectively. Second, *Permuted Trust* keeps the exact value distribution but reassigns it to the wrong cells; it under-performs CART on **24/24** dataset-seed combinations, isolating *correct cell assignment* – not merely a reasonable-looking gate – as the source of the gain. Third, CART beats the best *fixed* gate and *raw* input on **23/24** dataset-seed combinations each (the one tie/loss both on Diabetes seed 42). A dataset-level two-sided Wilcoxon signed-rank test ($n = 8$) confirms all three comparisons are significant: CART vs. Fixed gate, $+4.25$ pp (t-based 95% CI $[0.80, 7.71]$, $p = 0.0078$); vs. Permuted, $+4.06$ pp ($[1.61, 6.50]$, $p = 0.0078$); vs. Raw, $+5.95$ pp ($[1.34, 10.56]$, $p = 0.0078$).

Adaptive vs. the best possible fixed gate. A skeptical reader might ask whether some *other* fixed value than $r \equiv 0.8$ would close the gap. We select the best fixed gate $r^* \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ per dataset on a held-out validation split and compare it against CART on a disjoint test split. CART wins on **7 of 8 datasets**, losing only on Diabetes; the **mean** selected r^* **across seeds** varies substantially (0.40–0.87), showing no single global ratio suffices (full results in Appendix B, Table 6; per-seed results were not retained, see Limitations).

5.5 Generalization to Unseen Corruption

We evaluate CART and the five Group-A baselines under two unseen corruption types: unseen-magnitude noise ($\sim$1.5–2.5$\times$ the training scale) and single-feature distributional shift (full results

in Appendix B, Table 7). CART attains the best accuracy on **5/8** datasets under unseen-magnitude corruption and best/tied-best on **5/8** under feature shift.

5.6 Effect of the Corruption Training Schedule

Finally, we isolate the training-strategy from the architectural contribution by comparing CART (default sampled-rate schedule) against CART-fixed (fixed 20% rate). Averaged over 8 datasets, the sampled-rate schedule improves mean Acc@20 (79.5 vs. 79.4) and NR-AUC (78.8 vs. 78.4) with clean accuracy unchanged; results are directionally consistent but not uniform per dataset.

Taken together, these experiments indicate that CART's robustness is consistently tied to the reliability-gated use of repaired representations rather than to reconstruction or the corruption training schedule alone.

6 Discussion

Reliability is not reconstruction. TabAE-E2E (sharing CART's tokenizer) is a stronger reliability detector than a generic autoencoder, yet still not uniformly better than CART (Section 5.3), and *indiscriminate* use of a good reconstruction (Repair-only) is actively harmful (Section 5.4): reconstruction quality and reliability-guided *selective use* of it are separable capabilities, and the latter drives CART's advantage.

Reliability is not anomaly or importance. On held-out corrupted Adult data, high-SHAP (Lundberg & Lee, 2017) importance cells are corrupted at roughly the same rate as low-importance cells, and Isolation Forest (Liu et al., 2008) scores do not isolate the errors. In contrast, CART's trust clearly separates clean from corrupted cells (Figure 3), supporting our claim that anomaly $A(x_{ij})$, importance $I(x_{ij}, y)$, and reliability $T(x_{ij}, f)$ are distinct.

Failure cases. Diabetes is CART's most consistent weak point: narrowest trust separation, the only dataset where a fixed gate beats CART (Table 6), and where CART loses under unseen corruption (Table 7). It is also our smallest dataset – a genuine limitation, not an outlier to dismiss.

7 Limitations

Synthetic supervision. The trust head is directly supervised by synthetic corruption masks during training; "label-free" always means *without human-annotated real-world corruption*, and no mask is available at test time. **Corruption protocol.** Our standard mechanism (independent per-cell noise/replacement) does not cover contextual or neighbor-correlated corruption; the unseen-magnitude and feature-shift experiments are a step toward broader coverage, not a complete answer.

8 Conclusion

We introduced CART, which decomposes robust tabular prediction into a contextual, per-cell trust estimate and a true leave-one-out repair, adaptively fused before classification. The central finding is mechanistic rather than a leaderboard result: robust tabular prediction benefits from treating reliability as a *contextual, cell-level* decision – not a feature-level prior or a generic reconstruction objective – and a seven-way ablation, a fixed-gate comparison, and a statistical test all show the gain requires assigning the *correct* score to the *correct* cell, not merely having a repair mechanism or an arbitrary gate. This translates into stronger robustness across our benchmark suite (Table 1, Figure 2), reliability detection competitive with or better than reconstruction baselines sharing CART's tokenizer (Table 2), and an advantage that largely, though not universally, persists under unseen corruption (Table 7). The resulting perspective is that robust tabular prediction need not treat corruption as a record-level failure: it can instead identify which cells remain trustworthy and selectively intervene only where contextual evidence warrants repair.

AI USE STATEMENT

Generative AI tools (large language models) were used during the preparation of this manuscript to assist with literature discovery, debugging and editing of experimental code, and editing portions of the text for clarity and flow. AI-assisted code was reviewed and tested by the authors, and all reported experiments and numerical results were independently verified. No AI tools were used to generate data, prove mathematical claims, or produce the reported numerical results. The authors take responsibility for the final content of this work, including all code, claims, and results.

ETHICS STATEMENT

This work uses publicly available, de-identified tabular benchmark datasets (Adult, Bank, Churn, Credit, Diabetes, HELOC, Higgs, Magic) and does not involve human subjects research, new data collection, or personally identifiable information beyond what is already present in these standard benchmarks.

REPRODUCIBILITY STATEMENT

All datasets are publicly available benchmark tabular datasets (Section 4.1). The corruption protocols, baseline training procedures, and evaluation metrics are fully specified in Section 4; the model architecture and training objective are specified in Section 3. Exact per-dataset, per-seed numerical results underlying every table and figure in this paper are available in the accompanying result files, and the seven-way mechanistic ablation and statistical test are reproducible from a single saved model checkpoint per dataset. The implementation, pretrained checkpoints, experimental results, and evaluation scripts are available at the following anonymous repository: `https://anonymous.4open.science/r/cart_iclr-0FC3`.

REFERENCES

Sercan O. Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. SCARF: Self-supervised contrastive learning using random feature corruption. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):4308, 2014.

RK Bock et al. Magic telescope: a new ground-based gamma-ray observatory. *Nuclear Instruments and Methods in Physics Research Section A*, 516(2-3):511–528, 2004.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 785–794, 2016.

Y. Cheng et al. TabFSBench: Tabular benchmark for feature shifts in open environments. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

FICO. Explainable machine learning challenge. `https://community.fico.com/s/explainable-machine-learning-challenge`, 2017.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Rundong He and Jieming Shi. Prior-free tabular test-time adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=BgSDPE24pa.

Daniel Jarrett, Bogdan C. Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. HyperImpute: Generalized iterative imputation with automatic model selection. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

Kaggle. Bank customer churn prediction. https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 214–219, 1996.

Chorok Lee. Diagnosing conformal prediction failures under distribution shift: A covid-19 case study, 2026.

Hengyue Liang, Le Peng, and Ju Sun. Selective classification under distribution shifts, 2024.

Zhongyuan Liang, Zachary T. Rewolinski, Abhineet Agarwal, Tiffany M. Tang, and Bin Yu. Local MDI+: Local feature importances for tree-based models, 2026.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 413–422, 2008.

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

Alberto Archetti Mattia Sabella and Cinzia Cappiello Pietro Pinoli, Matteo Matteucci. QuAIL: Quality-aware inertial learning for robust training under data corruption. *arXiv preprint arXiv:2602.03686*, 2026.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 261, 1988.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1096–1103, 2008.

J. Wang, Z. Peng, X. Jia, Q. Lin, L. Ma, and B. Shi. Revisiting weakly supervised tabular anomaly detection from a cell-level perspective. *Neural Networks*, 203:109202, 2026. doi: 10.1016/j.neunet.2026.109202.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2): 2473–2480, 2009.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. VIME: Extending the success of self- and semi-supervised learning to tabular domain. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A Dataset Details

We evaluate CART on 8 standard tabular binary classification benchmarks, primarily sourced from the UCI Machine Learning Repository and OpenML (Vanschoren et al., 2013). The datasets vary significantly in scale, feature composition (mix of numerical and categorical), and domain.

- **Adult** (Kohavi, 1996): Predicts whether an individual's income exceeds $50K/year based on census demographic data. It contains a mix of categorical (e.g., education, occupation) and numerical (e.g., age, hours-per-week) features.
- **Bank** (Moro et al., 2014): Predicts whether a client will subscribe to a term deposit, based on direct marketing campaign data from a Portuguese banking institution.
- **Churn** (Kaggle, 2020): Predicts bank customer churn (whether a customer will close their account), containing features such as credit score, balance, and tenure.
- **Credit** (Yeh & Lien, 2009): Predicts default payments of credit card clients in Taiwan, utilizing features like payment history, bill statements, and demographic factors.
- **Diabetes** (Smith et al., 1988): Predicts the onset of diabetes within five years in patients of Pima Indian heritage, based on diagnostic medical measurements (purely numerical features).
- **HELOC** (FICO, 2017): Predicts whether a homeowner will repay their Home Equity Line of Credit (HELOC) account within two years, based on anonymized credit bureau data.
- **Higgs** (Baldi et al., 2014): A large-scale physics dataset to distinguish between a signal process which produces Higgs bosons and a background process, using purely numerical kinematic properties.
- **Magic** (Bock et al., 2004): Simulates registration of high-energy gamma particles in a ground-based atmospheric Cherenkov telescope, classifying signal (gamma) vs. background (hadron) using numerical image parameters.

B Full Per-Dataset Numerical Results

This appendix reports the full per-model, per-dataset breakdown (with standard deviations over 3 seeds) underlying the condensed tables and figures of the main text.

B.1 Overall Robust Prediction (Full Breakdown)

Table 4 is the full 6-model version of Table 1.

Table 4: Overall robust prediction performance under standard cell corruption. Accuracy (%, mean±std over 3 seeds) and NR-AUC (%); best value per dataset in **bold**, second-best underlined.

Dataset	Model	Clean Acc.	Acc@20%	NR-AUC
Adult	CatBoost	**83.89**±0.13	81.64±0.21	80.75±0.04
	XGBoost	83.85±0.11	81.07±0.17	80.17±0.09
	FT-Transf.	83.63±0.06	81.53±0.36	80.52±0.19
	MLP	82.36±0.50	78.84±0.50	77.71±0.20
	QuAIL	82.45±0.34	79.25±0.17	78.39±0.15
	CART (Ours)	83.76±0.10	**82.03**±0.29	**80.98**±0.06
Bank	CatBoost	89.74±0.15	**89.01**±0.24	88.47±0.15
	XGBoost	**89.88**±0.12	88.78±0.10	88.24±0.23
	FT-Transf.	89.58±0.23	**89.01**±0.15	88.40±0.09
	MLP	89.15±0.13	87.25±0.47	86.47±0.61
	QuAIL	89.60±0.03	88.15±0.35	87.61±0.31
	CART (Ours)	89.68±0.14	88.94±0.03	**88.66**±0.04
Churn	CatBoost	92.80±0.23	89.51±0.62	88.44±0.12
	XGBoost	**93.11**±0.15	89.07±0.23	88.45±0.17
	FT-Transf.	91.69±0.54	89.02±1.81	88.26±1.50
	MLP	91.69±0.47	86.67±0.74	85.94±0.66
	QuAIL	92.44±1.32	88.84±0.28	87.37±0.66
	CART (Ours)	92.76±0.31	**90.40**±0.83	**89.90**±0.35
Credit	CatBoost	81.65±0.20	81.05±0.19	**80.80**±0.17
	XGBoost	81.72±0.13	80.71±0.35	80.37±0.25
	FT-Transf.	**81.78**±0.04	80.64±0.08	80.36±0.20
	MLP	80.49±0.60	79.03±0.38	78.43±0.28
	QuAIL	81.62±0.16	80.47±0.16	79.98±0.12
	CART (Ours)	81.74±0.15	**81.09**±0.16	80.71±0.07
Diabetes	CatBoost	73.85±1.32	71.84±4.07	70.20±1.08
	XGBoost	72.13±1.32	69.25±2.63	68.42±1.03
	FT-Transf.	75.29±1.32	**74.14**±1.49	71.01±0.28
	MLP	74.14±2.99	72.41±4.80	69.25±2.59
	QuAIL	75.57±2.77	73.28±3.45	71.78±0.70
	CART (Ours)	**77.30**±3.26	**74.14**±2.59	**71.84**±1.97
HELOC	CatBoost	**72.87**±0.31	70.47±0.77	69.94±0.32
	XGBoost	72.60±0.35	69.44±0.81	69.05±0.35
	FT-Transf.	72.60±0.41	**71.47**±0.44	70.10±0.67
	MLP	69.29±1.30	65.09±1.59	65.19±0.68
	QuAIL	71.89±0.71	69.36±0.54	68.16±0.45
	CART (Ours)	72.44±0.41	71.09±0.47	**70.39**±0.16
Higgs	CatBoost	71.12±0.40	67.25±0.32	66.30±0.15
	XGBoost	70.78±0.21	66.61±0.79	65.63±0.16
	FT-Transf.	70.51±0.17	66.83±0.21	65.66±0.07
	MLP	67.65±0.31	62.68±0.59	61.92±0.21
	QuAIL	69.96±0.31	65.84±0.29	64.80±0.44
	CART (Ours)	**71.34**±0.26	**67.97**±0.26	**66.90**±0.18
Magic	CatBoost	86.11±0.19	81.70±0.44	80.08±0.31
	XGBoost	86.13±0.12	81.83±0.56	79.97±0.21
	FT-Transf.	85.54±0.58	81.21±0.11	79.55±0.16
	MLP	85.62±0.26	80.39±0.62	78.67±0.28
	QuAIL	85.42±0.28	80.98±0.44	78.96±0.24
	CART (Ours)	**86.46**±0.74	**82.86**±0.27	**80.60**±0.17

B.2 Cell Reliability and Reconstruction (Full Breakdown)

Table 5 is the full per-dataset version of Table 2; Figure 5 visualizes the dataset-averaged metrics.

Table 5: Cell-level reliability detection and recovery (mean±std over 3 seeds). Cell-AUROC/AUPRC (↑), numerical reconstruction MAE (↓), categorical reconstruction accuracy (%, ↑); best value per dataset in **bold**, second-best underlined; "–" = dataset has no categorical features.

Dataset	Model	Cell-AUROC	Cell-AUPRC	Num. MAE	Cat. Acc.
Adult	CART (Ours)	**0.926**±0.000	**0.979**±0.000	**0.536**±0.009	**70.818**±0.425
	DAE	0.919±0.001	0.976±0.000	0.652±0.003	48.145±0.969
	TabAE	0.920±0.000	0.977±0.000	0.676±0.020	50.001±0.310
	VAE	0.904±0.001	0.971±0.001	0.614±0.004	54.458±0.562
Bank	CART (Ours)	**0.901**±0.001	**0.970**±0.000	0.531±0.005	**69.993**±0.695
	DAE	0.869±0.004	0.956±0.001	0.642±0.007	43.150±2.203
	TabAE	0.900±0.001	0.969±0.000	0.619±0.006	43.501±0.912
	VAE	0.828±0.003	0.943±0.001	**0.486**±0.004	47.957±1.071
Churn	CART (Ours)	0.788±0.009	**0.925**±0.005	**0.605**±0.015	**68.400**±1.417
	DAE	**0.792**±0.006	0.921±0.004	0.793±0.009	46.001±2.172
	TabAE	0.774±0.017	0.917±0.007	0.871±0.042	49.754±0.983
	VAE	0.747±0.010	0.905±0.004	0.733±0.015	49.043±0.133
Credit	CART (Ours)	**0.963**±0.004	**0.988**±0.001	**0.396**±0.006	–
	DAE	0.880±0.003	0.951±0.001	0.474±0.005	–
	TabAE	0.951±0.001	0.984±0.000	0.453±0.008	–
	VAE	0.850±0.002	0.946±0.001	0.421±0.005	–
Diabetes	CART (Ours)	0.724±0.049	0.888±0.029	0.832±0.084	–
	DAE	**0.769**±0.008	**0.905**±0.008	0.834±0.006	–
	TabAE	0.718±0.031	0.876±0.019	1.002±0.021	–
	VAE	0.764±0.018	0.898±0.010	**0.723**±0.050	–
HELOC	CART (Ours)	**0.884**±0.008	**0.959**±0.003	**0.430**±0.004	–
	DAE	0.858±0.004	0.941±0.001	0.488±0.007	–
	TabAE	0.874±0.003	0.952±0.000	0.558±0.019	–
	VAE	0.824±0.006	0.929±0.003	0.482±0.005	–
Higgs	CART (Ours)	0.844±0.001	0.947±0.001	**0.648**±0.003	–
	DAE	0.786±0.002	0.907±0.001	0.700±0.004	–
	TabAE	**0.850**±0.001	**0.949**±0.000	0.722±0.005	–
	VAE	0.745±0.002	0.894±0.001	0.736±0.001	–
Magic	CART (Ours)	0.876±0.002	**0.954**±0.000	0.474±0.004	–
	DAE	0.847±0.001	0.934±0.001	0.510±0.003	–
	TabAE	**0.876**±0.004	0.953±0.002	0.522±0.008	–
	VAE	0.833±0.005	0.932±0.001	**0.453**±0.006	–

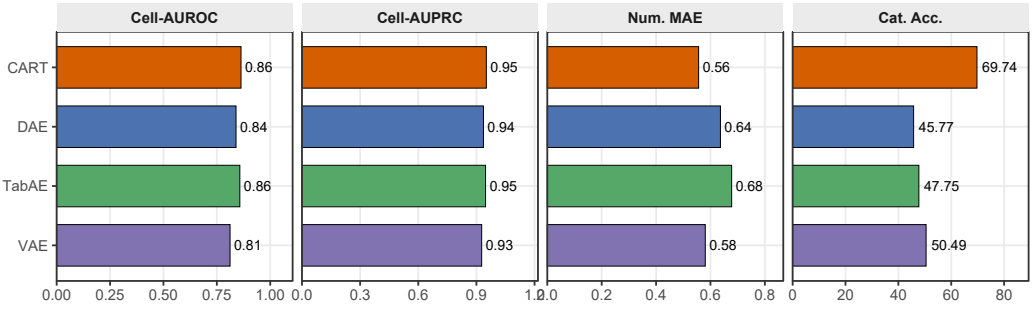

Figure 5: CART vs. reconstruction-oriented baselines that share its tokenizer (DAE, TabAE-E2E, VAE). Cell-AUROC, Cell-AUPRC, and numerical MAE are averaged over all 8 datasets; categorical reconstruction accuracy is averaged only over the 3 datasets (Adult, Bank, Churn) that contain categorical features. CART is competitive with or ahead of all three baselines on reliability detection and categorical recovery, while TabAE-E2E – the no-LOO ablation – is its closest competitor.

B.3 Adaptive vs. Best Fixed Gate (Full Breakdown)

Table 6 reports the per-dataset validation-selected fixed gate r^* and its test accuracy against CART, referenced from Section 5.4.

Note on evaluation protocol: For this specific comparison, we split the 20%-corruption test set into disjoint validation and test halves to select r^* without data leakage. The selected r^* and test accuracy are averaged over 3 seeds (per-seed standard deviations were not retained for this specific sweep).

Table 6: Adaptive (CART) vs. the best fixed gate selected on a held-out validation split, then evaluated on a disjoint held-out test split. Accuracy (%) at 20% corruption; best value per row in **bold**.

Dataset	**Mean Best r^* (val)**	**Fixed gate (test)**	**CART (test)**
Adult	0.60	81.12	**82.30**
Bank	0.60	88.15	**88.73**
Churn	0.53	87.82	**90.76**
Credit	0.40	79.56	**81.24**
Diabetes	0.60	**71.26**	69.54
HELOC	0.87	67.82	**71.24**
Higgs	0.73	60.68	**68.12**
Magic	0.80	70.08	**82.25**

B.4 Generalization to Unseen Corruption (Full Breakdown)

Table 7 is the full per-dataset version of the summary given in Section 5.5.

Table 7: Robustness under corruption types not used during training. Accuracy (%), averaged over 3 seeds; "Best other" is the best of the five Group-A baselines; **bold** marks whichever of the two is higher.

	Unseen magnitude		**Feature shift**	
Dataset	**Best other**	**CART**	**Best other**	**CART**
Adult	**83.61**	83.48	83.43	**83.66**
Bank	88.92	**89.48**	**89.75**	89.45
Churn	**90.40**	89.60	**91.82**	**91.82**
Credit	79.64	**81.10**	81.66	**81.75**
Diabetes	**70.69**	70.40	**75.29**	74.71
HELOC	68.87	**71.00**	**72.62**	72.22
Higgs	65.26	**67.49**	67.92	**67.95**
Magic	79.97	**81.53**	85.82	**86.14**