"WHERE CAN I TRUST YOU?": BOUNDARY-AWARE EVALUATION OF SURROGATE FIDELITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Surrogate models are commonly evaluated by how often they agree with their teacher model over an evaluation set. Local variation in this agreement is well known, but its structure and consequences are less clear. We ask whether disagreement is systematically concentrated near the teacher's decision boundary and whether retaining that structure provides information beyond a single global score. Across several datasets and surrogate model classes, we find substantially lower fidelity near teacher decision boundaries under two different methods of identifying near-boundary examples. Moreover, conditioning agreement on confidence-defined regions improves prediction of teacher–surrogate agreement when evaluation-set composition changes, relative to the global score alone. Yet surrogates that agree equally well with the teacher both globally and near the decision boundary can respond very differently to changes selected using the surrogate itself. Finally, we show that independently trained deep teachers can agree on most predictions while identifying different examples as lying near their decision boundaries, complicating the use of those boundaries as stable reference regions for evaluating surrogates. Together, these results show that surrogate fidelity depends not only on how often a surrogate agrees with its teacher, but also on where that agreement holds and, for deep models, how stable the teacher's decision boundary is across training runs.

1 INTRODUCTION

Surrogate models are only as useful as their ability to reliably stand in for the models they approximate. This reliability is typically evaluated through fidelity: the agreement between a surrogate and its teacher over an evaluation set. Prior work has shown that this agreement can vary substantially across the evaluation space. What remains less clear is whether that variation has systematic structure and whether retaining that structure provides information beyond a single global score.

In this work, we study the structure of teacher–surrogate disagreement for a fixed pair, with particular attention to the teacher's decision boundary. We ask whether disagreement is systematically concentrated near the teacher's decision boundary, whether similar levels of global and boundary-local fidelity correspond to similar counterfactual transfer rates, and whether measuring agreement across teacher-confidence levels better predicts fidelity under changes in the evaluation composition. We further examine whether the boundary regions used for this analysis are stable across independently trained deep teachers.

We characterize fidelity as a conditional property of a teacher–surrogate pair over an evaluation space, rather than as a single global quantity. Our contributions are: (1) characterizing how teacher–surrogate disagreement varies with proximity to the teacher's decision boundary, using confidence-based and geometric proxies; (2) evaluating whether confidence-stratified profiles predict fidelity better than a global score when the evaluation composition changes; (3) examining what matched global and boundary-local fidelity does and does not establish about example-based counterfactual transfer; and (4) studying whether independently trained deep teachers identify the same examples as lying near their decision boundaries.

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.

2 Related Work

Fidelity arises from a broader problem of determining how well one model preserves the behavior of another. In knowledge distillation, compact student models are trained to reproduce the outputs or intermediate representations of larger teacher models (Hinton et al., 2015; Romero et al., 2015). Surrogate-based interpretability uses a teacher–student setup for a different purpose: a simpler, more interpretable model is trained to approximate the behavior of an otherwise difficult-to-interpret predictor (Ribeiro et al., 2016; Lyu et al., 2024).

The quality of this approximation is commonly evaluated through fidelity, reported as an aggregate measure of agreement between a surrogate and its teacher over an evaluation set (Setzu et al., 2021; Confalonieri et al., 2021; Henckaerts et al., 2022; Mariotti et al., 2023). Such aggregation can obscure heterogeneity in approximation quality. Balagopalan et al. (2022), for example, show that both local and global explanation models can exhibit substantially different fidelity across protected subgroups, so that average fidelity may overstate explanation quality for some populations. Their analysis conditions fidelity on externally defined group membership. In contrast, we define regions using the teacher's own decision structure and ask whether disagreement is systematically concentrated near its decision boundary.

Locality also plays a central role in how surrogate explanations are constructed and evaluated. Ribeiro et al. (2016) fit interpretable surrogates in neighborhoods around individual predictions. Laugel et al. (2018) show that the choice of neighborhood is important for accurately approximating local decision-boundary structure, while Poyiadzi et al. (2021) subsequently characterize locality as a continuum of coverage, studying how the fidelity–complexity tradeoff changes as the region used to fit the surrogate contracts. More recently, Huynh et al. (2026) distinguish local from global validation contexts, showing that the measured fidelity of an individual rule explanation can decrease when the same rule is evaluated beyond its original neighborhood.

Related work has also questioned what high predictive fidelity actually establishes about a surrogate. Mariotti et al. (2023) show that models with high prediction-level fidelity can nevertheless differ substantially in feature-attribution structure, motivating ShapGAP as a comparison of model explanations rather than outputs alone. Moreover, Eshbaugh (2026) distinguishes fidelity to a learned model from recovery of the task-relevant structure underlying predictive performance, showing that a surrogate can closely approximate a neural network's learned function while failing to preserve its predictive advantage. The present work addresses a complementary limitation: for a fixed teacher–surrogate pair, we ask whether predictive disagreement is systematically organized relative to the teacher's decision boundary and whether retaining that structure improves prediction of fidelity when evaluation composition changes. We further examine whether matching both global and boundary-local agreement is sufficient to produce similar counterfactual transfer behavior.

Prior work has examined the reproducibility of neural-network decision boundaries across independent training runs. Somepalli et al. (2022) find substantial similarity in decision regions across random initializations of the same architecture, particularly for wide convolutional networks, while also observing systematic differences across model families. Lei et al. (2025) study decision-boundary variability across training repeats, operationalizing it through predictive disagreement between independently trained networks and relating lower variability to improved generalization. Our analysis concerns a different notion of boundary stability: whether independently trained teachers identify the same evaluation examples as lying near their decision boundaries.

These studies motivate our focus on how teacher–surrogate disagreement is distributed relative to the teacher's decision boundary, whether that structure improves prediction of fidelity when evaluation composition changes, what matched global and boundary-local agreement establishes about counterfactual transfer, and how stable the relevant boundary regions are across independently trained teachers.

3 Methodology

Our methodology is designed to characterize teacher–surrogate agreement beyond a single global fidelity score. We first define global and boundary-conditioned fidelity under two notions of boundary proximity, then describe protocols for predicting fidelity under changes in evaluation composi-

tion and for measuring example-based counterfactual transfer. Finally, we describe our analysis of boundary stability across independently trained deep teachers.

3.1 TEACHER–SURROGATE FIDELITY

Let T represent a teacher model and S_T a surrogate model trained on input–output pairs $(x, T(x))$ to reproduce the teacher's predicted class rather than the ground-truth label. Fidelity is measured over an evaluation set $X = \{x_i\}_{i=1}^n$ and quantifies agreement with the teacher rather than accuracy against ground-truth labels. Global fidelity is defined in Equation 1. This scalar gives the fraction of evaluation examples on which the surrogate and teacher make the same prediction.

$$F(T, S_T; X) = \frac{1}{|X|} \sum_{x \in X} \mathbf{1}[T(x) = S_T(x)] \tag{1}$$

To examine this agreement, our tabular experiments use a common teacher–surrogate training setup. Features are standardized using statistics estimated from the training partition only. Teachers are binary multilayer perceptrons with hidden widths 64 and 32 and ReLU activations, trained on ground-truth labels using binary cross-entropy. Our primary capacity analyses use decision-tree surrogates; in Section 4.1, we additionally evaluate logistic-regression, k-nearest-neighbor, and multilayer-perceptron surrogates to test whether boundary-local disagreement is specific to the decision-tree model class. Experiments are repeated across ten random seeds unless otherwise stated. Full dataset, preprocessing, model, training, and boundary-computation details are provided in Appendix A.

3.2 BOUNDARY-AWARE FIDELITY

Global fidelity does not distinguish between disagreement occurring in different regions of the evaluation space. We therefore condition fidelity on proximity to the teacher's decision boundary.

Confidence-based proximity. For a binary teacher, let $f_T(x) \in \mathbb{R}$ denote the teacher logit, such that $T(x) = \mathbf{1}[f_T(x) \geq 0]$.

Our first measure uses the magnitude of the teacher logit, $c_T(x) = |f_T(x)|$. Smaller values indicate lower teacher confidence and serve as a proxy for greater proximity to the teacher's prediction transition. For a boundary fraction $\alpha \in (0, 1]$, we define $B_\alpha^{\text{conf}}(T; X)$ as the $\lceil \alpha |X| \rceil$ evaluation examples with the smallest values of $c_T(x)$.

Segment-crossing proximity. Because confidence-based proximity depends directly on the teacher's output score, we also use a geometric proxy that depends only on teacher predictions and the geometry of the evaluation set. For each $x \in X$, we identify the nearest observed evaluation example $x' \in X$ for which $T(x') \neq T(x)$, using Euclidean distance in the standardized feature space. We then perform a binary search along the line segment connecting x and x' to approximate a point at which the teacher prediction changes. The distance from x to this estimated crossing defines $d_T^{\text{seg}}(x)$. This quantity is an approximate segment-crossing distance and should not be interpreted as the exact minimum Euclidean distance to the teacher's decision boundary. We define $B_\alpha^{\text{seg}}(T; X)$ as the $\lceil \alpha |X| \rceil$ examples with the smallest values of $d_T^{\text{seg}}(x)$.

For either boundary definition B_α, boundary-conditioned fidelity is given in Equation 2. Equation 3 reports the corresponding boundary gap where positive values indicate lower fidelity on the selected examples than on the evaluation set as a whole.

$$F_\alpha(T, S_T; X) = \frac{1}{|B_\alpha|} \sum_{x \in B_\alpha} \mathbf{1}[T(x) = S_T(x)] \tag{2}$$

$$\Delta_\alpha = F(T, S_T; X) - F_\alpha(T, S_T; X) \tag{3}$$

We use these measures to examine how disagreement near the boundary varies with region size, surrogate capacity, and surrogate model class. For the boundary-local analysis, we evaluate boundary fractions $\alpha \in \{0.10, 0.20, 0.25, 0.40\}$. Our primary capacity analysis uses decision-tree surrogates

with maximum depths $\{1, 2, 3, 5, 7, 10\}$. To test whether the observed boundary-local structure is specific to decision trees, we additionally compare representative decision-tree, logistic-regression, k-nearest-neighbor, and multilayer-perceptron surrogates at $\alpha = 0.25$. These configurations are a depth-5 decision tree, logistic regression with $C = 1$, k-NN with $k = 5$, and a one-hidden-layer MLP with width 32.

3.3 Predicting Fidelity Under Changes in Evaluation Composition

Having defined fidelity within selected regions, we examine whether measuring agreement across teacher-confidence levels improves predictions of fidelity when evaluation composition changes. For each tabular dataset, we reserve 60% of the data for training and 40% for evaluation. The evaluation partition is then divided evenly into disjoint source and target pools, stratified by the teacher's prediction. The source pool is used to estimate agreement within confidence bins; the target pool provides the bin weights used to predict target fidelity and the observed agreement used to evaluate those predictions.

We partition the source pool into three quantile bins according to teacher confidence $c_T(x) = |f_T(x)|$. For each bin b, let X_b denote the source-pool examples assigned to that bin. We estimate conditional fidelity as $F_b = \frac{1}{|X_b|} \sum_{x \in X_b} \mathbf{1}[T(x) = S_T(x)]$. The source-derived bin boundaries are then applied unchanged to the target examples. Let $w_b = \frac{|X_{\text{target},b}|}{|X_{\text{target}}|}$ denote the proportion of target examples assigned to bin b.

We predict target fidelity as

$$\widehat{F}_{\text{profile}} = \sum_b w_b F_b. \tag{4}$$

This prediction weights source-bin agreement rates by the composition of the target pool. We compare it with a baseline using source-pool global fidelity directly, $\widehat{F}_{\text{global}} = F(T, S_T; X_{\text{source}})$.

To compare these predictors under changes in evaluation composition, we divide the target pool at its median teacher confidence into lower- and higher-confidence subsets. We treat each subset as a separate target distribution and apply both fidelity predictors to each. Prediction error is measured as the absolute difference between predicted and observed target fidelity. Our primary analysis uses three confidence bins; we additionally evaluate five bins as a sensitivity analysis.

3.4 Example-Based Counterfactual Transfer

Beyond predicting fidelity under changes in evaluation composition, we examine whether teacher–surrogate agreement corresponds to similar behavior under changes selected using the surrogate. We use an example-based counterfactual transfer protocol with disjoint training, candidate, and test partitions containing 60%, 20%, and 20% of the data, respectively.

For a test example x, we consider only cases in which the teacher and surrogate initially agree, $T(x) = S_T(x)$. Let y' denote the class opposite to the surrogate's prediction on x. From the candidate set $X_{\text{candidate}}$, we select the nearest example

$$x^\star = \underset{z \in X_{\text{candidate}}}{\arg\min} \|x - z\|_2 \quad \text{subject to} \quad S_T(z) = y', \tag{5}$$

where distance is measured in the standardized feature space. The selected candidate therefore represents the nearest available example that changes the surrogate's prediction. Counterfactual transfer occurs when the teacher also assigns that target class to the selected candidate, $T(x^\star) = y'$.

We define the transfer rate as the fraction of eligible test examples for which this condition holds. This protocol measures whether a surrogate-selected change that alters the surrogate's prediction also induces the corresponding change in the teacher; it does not assume that the selected example is a causal or actionable counterfactual.

To examine whether observational fidelity is sufficient to characterize transfer behavior, we compare surrogate configurations with similar global fidelity and boundary-conditioned fidelity at $\alpha = 0.25$.

3.5 Boundary Stability Across Deep Teachers

Boundary-aware fidelity is defined relative to a teacher, but in deep learning the teacher itself may vary across independent training runs. We therefore measure whether independently trained teachers identify the same examples as lying near their decision boundaries.

For a multiclass teacher T, let $f_T(x) \in \mathbb{R}^K$ denote its logit vector. Let $f_T^{(1)}(x)$ and $f_T^{(2)}(x)$ denote its largest and second-largest logits, respectively. We define the teacher's prediction margin as

$$m_T(x) = f_T^{(1)}(x) - f_T^{(2)}(x). \tag{6}$$

Smaller margins indicate greater proximity to a prediction transition.

Because absolute logit scales are not necessarily comparable across independently trained networks, we define boundary regions by rank rather than by a common margin threshold. For a boundary fraction α, the set $B_\alpha(T; X)$ contains the $\lceil \alpha |X| \rceil$ examples with the smallest top-two logit margins for teacher T.

We train five independently trained teachers on MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) and compare every unique teacher pair.

For teachers T_i and T_j, boundary-set stability is measured using Jaccard overlap (Equation 7).

$$J_\alpha(T_i, T_j) = \frac{|B_\alpha(T_i; X) \cap B_\alpha(T_j; X)|}{|B_\alpha(T_i; X) \cup B_\alpha(T_j; X)|}. \tag{7}$$

We additionally measure pairwise predictive agreement globally and over the union and intersection of the two teachers' boundary regions. These comparisons are reported descriptively across teacher pairs, because individual teachers participate in multiple pairs so the pairwise observations are not independent.

4 Results

Using the methods described in Section 3, we examine where teacher-surrogate disagreement occurs, whether confidence-stratified fidelity improves prediction when evaluation composition changes, and what global and boundary-local agreement establish about counterfactual transfer. We then assess how stable the boundary regions are across independently trained teachers.

4.1 Boundary-Local Structure of Surrogate Disagreement

To examine whether disagreement is concentrated near the teacher's decision boundary, we compare global and boundary-local fidelity across a range of decision-tree surrogate capacities. Boundary-local fidelity is consistently lower than global fidelity on three datasets—Breast Cancer (Wolberg et al., 1993), Diabetes (Efron et al., 2004), and Wine (Aeberhard & Forina, 1992)—while Iris (Fisher, 1936) exhibits little difference between the two. At a boundary fraction of 25%, global fidelity exceeds confidence-based boundary fidelity by about 12, 19, and 10 percentage points (pp) on Breast Cancer, Diabetes, and Wine, respectively, compared with 1.3 pp on Iris.

Figure 1 shows that this gap widens as evaluation is restricted to examples closer to the teacher's decision boundary. The same qualitative pattern appears under both the confidence-based and segment-crossing measures, suggesting that the observed concentration of disagreement is not specific to a single way of identifying examples near the boundary.

We next test whether this pattern is specific to decision-tree surrogates. Table 1 compares representative decision-tree, logistic-regression, k-nearest-neighbor, and MLP surrogates at a fixed boundary fraction of 25%. On Breast Cancer, mean boundary-local fidelity is lower than global fidelity for all four surrogate families under both measures, with smaller gaps for the higher-fidelity logistic-regression and MLP surrogates. The result is particularly pronounced on Diabetes: confidence-based boundary fidelity is lower than global fidelity by approximately 19–22 pp across all four surrogate families, with a positive gap in every random seed.

Wine provides an informative contrast. The decision-tree surrogate exhibits a substantial boundary-local gap, while logistic regression, k-NN, and the MLP achieve approximately 98.5%–99.6%

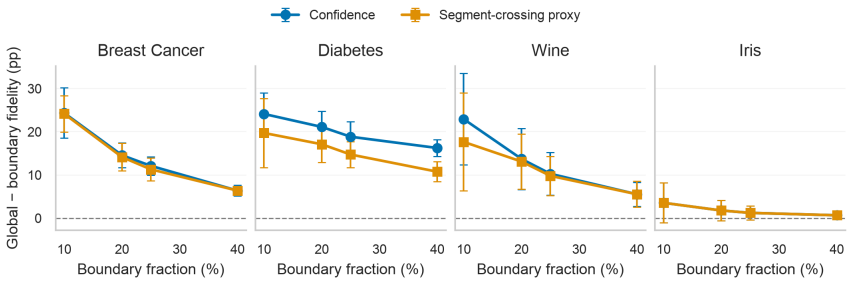

Figure 1: Global-to-boundary fidelity gaps, in percentage points, as the boundary fraction varies under confidence-based and approximate segment-crossing measures. Points show mean gaps across random seeds after averaging over decision-tree depths within each seed; error bars show approximate 95% confidence intervals across seeds.

Dataset	Surrogate	Global fidelity (%)	Confidence-based boundary Gap (pp) mean ± 95% CI	Confidence-based boundary Pos.	Segment-crossing boundary Gap (pp) mean ± 95% CI	Segment-crossing boundary Pos.
Breast Cancer	Decision tree	94.0	11.9 ± 2.2	10/10	10.1 ± 2.8	10/10
	Logistic regression	98.2	5.2 ± 1.6	10/10	5.2 ± 1.6	10/10
	k-NN	96.4	10.6 ± 3.4	10/10	9.9 ± 3.6	10/10
	MLP	98.8	3.4 ± 1.7	8/10	3.2 ± 1.6	8/10
Diabetes	Decision tree	73.6	18.6 ± 4.2	10/10	13.9 ± 4.6	9/10
	Logistic regression	79.5	20.7 ± 3.4	10/10	16.2 ± 3.7	10/10
	k-NN	80.7	20.4 ± 4.7	10/10	15.7 ± 4.3	10/10
	MLP	82.9	22.3 ± 4.1	10/10	13.8 ± 3.4	10/10
Wine	Decision tree	95.0	11.4 ± 5.2	9/10	11.4 ± 4.4	9/10
	Logistic regression	99.4	1.6 ± 2.2	2/10	1.6 ± 2.2	2/10
	k-NN	98.5	3.5 ± 3.1	5/10	3.5 ± 3.1	5/10
	MLP	99.6	1.1 ± 1.4	2/10	0.3 ± 1.1	1/10
Iris	Decision tree	99.6	1.2 ± 1.6	2/10	1.2 ± 1.6	2/10
	Logistic regression	99.6	1.2 ± 1.6	2/10	1.2 ± 1.6	2/10
	k-NN	99.3	1.8 ± 1.8	3/10	1.8 ± 1.8	3/10
	MLP	99.6	1.2 ± 1.6	2/10	1.2 ± 1.6	2/10

Table 1: Boundary-local fidelity degradation across surrogate families at $\alpha = 0.25$. Global fidelity and boundary gaps are averaged over ten seeds; gaps are reported in percentage points with approximate 95% confidence intervals. "Pos." counts seeds for which global fidelity exceeds boundary-local fidelity.

global fidelity and show much smaller degradation. Thus, the boundary-local concentration of disagreement is not an artifact of decision-tree surrogates, although its magnitude is smaller when the surrogate reproduces the teacher almost everywhere. More broadly, globally aggregated fidelity can conceal systematic spatial structure in teacher–surrogate disagreement across substantially different surrogate model classes. Additional decision-tree capacity results and surrogate-family robustness analyses are reported in Appendix B.

4.2 Predicting Fidelity Under Changes in Evaluation Composition

The concentration of disagreement near the teacher's decision boundary suggests that conditioning on teacher confidence may improve prediction of surrogate fidelity when evaluation composition changes. We compare a global-fidelity baseline with a confidence-stratified profile that predicts fidelity on the target set based on its composition across confidence regions. Across 180 evaluation conditions with altered composition, the global baseline yields a mean absolute error of 7.67 pp, compared with 4.62 pp for the three-bin stratified profile, a 39.7% reduction in error.

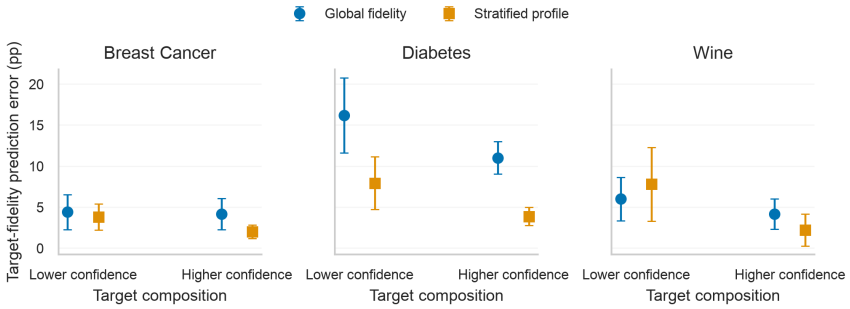

Figure 2: Confidence-stratified fidelity improves prediction when evaluation composition changes. Mean absolute error in predicting target-set fidelity under shifts toward lower- and higher-confidence regions. Predictions using a three-bin confidence-stratified fidelity profile are compared with a single global-fidelity baseline. Error bars indicate 95% confidence intervals across random seeds, after averaging over surrogate depths within each seed.

Figure 2 shows that the improvement is especially pronounced under shifts toward higher-confidence regions, while gains under shifts toward lower-confidence regions are more modest. The stratified profile outperforms the global baseline in 121 of 180 conditions, with 9 ties. At the dataset level, mean absolute error decreases from 4.30 pp to 2.93 pp on Breast Cancer and from 13.62 pp to 5.92 pp on Diabetes, while remaining nearly unchanged on Wine (5.09 pp versus 5.02 pp). A complementary analysis across representative decision-tree, logistic-regression, k-NN, and MLP surrogates yields the same overall pattern: mean absolute prediction error decreases from 6.79 pp to 3.72 pp, with lower mean error for the stratified profile in each surrogate family. These results show that agreement measured across teacher-confidence regions provides predictive information about target fidelity beyond the source-set global score when evaluation composition changes. Additional sensitivity analyses for the number of confidence bins and surrogate-family robustness are reported in Appendix C.

4.3 Counterfactual Transfer Under Matched Fidelity

The predictive value of boundary-aware fidelity does not mean that observational agreement is sufficient to characterize downstream behavior. We therefore evaluate example-based counterfactual transfer: whether a counterfactual selected using the surrogate induces the same target-class change in the teacher. On Breast Cancer, two surrogates can have identical global fidelity (94.74%) and identical boundary-conditioned fidelity (82.76%) while exhibiting sharply different counterfactual transfer rates: 73.15% versus 24.07%. Thus, matching both global and boundary-local observational agreement does not guarantee matched behavior under surrogate-selected changes.

Across datasets, surrogate configurations with similar global fidelity can exhibit substantial variation in counterfactual transfer, producing vertical spread in Figure 3. The highlighted Breast Cancer pair provides the clearest controlled example because the two surrogates are matched on both global and boundary-conditioned fidelity. These results show that observational agreement, even when localized, does not fully characterize how a surrogate will behave under changes selected using that surrogate. Additional results across surrogate capacities and matched-fidelity surrogate pairs are reported in Appendix D.

4.4 Boundary Stability Across Deep Teachers

Boundary-aware evaluation treats the teacher's decision boundary as a meaningful reference structure. In deep networks, however, independently trained teachers may make nearly identical predictions while identifying different examples as being near the decision boundary. We observe this behavior on both MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009). At a boundary fraction of 10%, independently trained MNIST teachers achieve approximately 99.2% pairwise predictive agreement, yet the sets of examples they identify as near the decision boundary have a mean Jaccard overlap of only 39.4%. For CIFAR-10, pairwise agreement remains approximately 96.0%, while mean boundary-set overlap is only 44.3%.

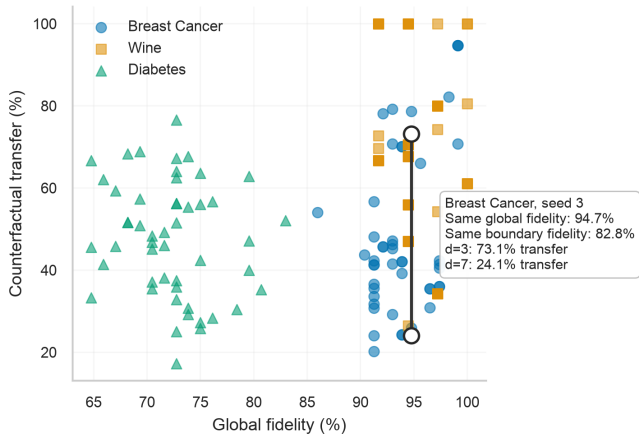

Figure 3: Similar observational agreement can correspond to substantial differences in counterfactual transfer. Each point represents one surrogate configuration, with global fidelity on the x-axis and counterfactual transfer on the y-axis. The vertical spread at similar global fidelity indicates that aggregate agreement does not uniquely determine counterfactual behavior. The highlighted Breast Cancer pair has identical global and boundary-local fidelity, but counterfactual transfer rates of 73.1% and 24.1%.

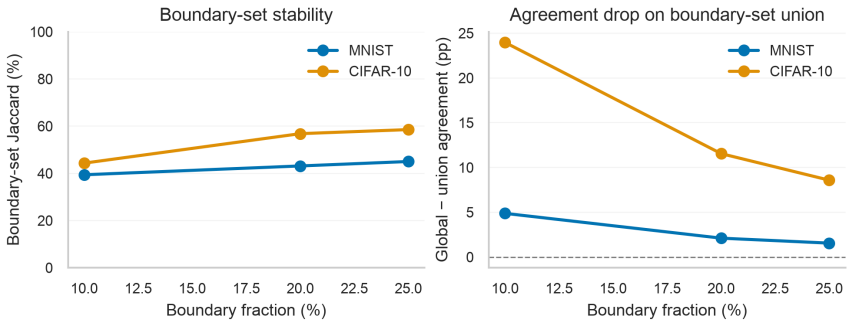

Figure 4: Boundary regions are only partially stable across independently trained deep teachers. Left: mean Jaccard overlap between teachers' lowest-margin examples as the boundary fraction varies. Right: the drop from global pairwise predictive agreement to agreement over the union of the two teachers' boundary regions. Results are shown for independently trained teachers on MNIST and CIFAR-10.

Figure 4 further shows that this partial boundary-set overlap is accompanied by lower predictive agreement within those regions. On CIFAR-10, predictive agreement drops from approximately 96.0% globally to 72.1% over the union of the two teachers' tightest boundary regions, a decrease of roughly 24 percentage points. The corresponding drop on MNIST is smaller but remains concentrated near the boundary. These results suggest that boundary-aware fidelity in deep networks should be interpreted relative to the particular teacher being evaluated unless the relevant boundary regions are stable across independent training runs. Individual teacher accuracies and additional pairwise boundary-set stability statistics are reported in Appendix E.

5 Discussion

Teacher–surrogate disagreement is systematically organized relative to the teacher's decision boundary, and retaining this structure improves prediction of surrogate fidelity when evaluation composition changes. This structure is therefore not merely descriptive; it provides information that is useful for anticipating agreement under changes in the evaluation set. Fidelity is better understood as conditional on regions of the evaluation space, not only as a scalar average.

However, localizing fidelity improves characterization without changing what fidelity can establish—it remains a measure of observational agreement. The matched Breast Cancer surrogates make this limitation concrete: identical global and boundary-conditioned fidelity can coexist with sharply different counterfactual transfer behavior. This distinction is important because a surrogate may reproduce the teacher's observed predictions while responding differently to changes selected using the surrogate itself. Boundary-aware fidelity should therefore be interpreted as additional evidence about the teacher–surrogate relationship rather than as a guarantee that the surrogate will preserve every downstream behavior of interest. Which forms of agreement matter will ultimately depend on the intended use of the surrogate.

A further complication is that the regions used for boundary-aware evaluation may themselves depend on the teacher being evaluated. Our deep-network experiments show that independently trained teachers can exhibit very high global predictive agreement while identifying substantially different examples as belonging to their boundary regions. Thus, "near the decision boundary" should not necessarily be treated as a model-independent property of the evaluation space. Boundary-aware fidelity in deep networks is therefore best interpreted relative to a particular teacher unless boundary stability across training runs has been established.

These conclusions should be interpreted within the scope of our experiments. In particular:

- **Empirical scope.** Our experiments cover a limited set of datasets and model families. The tabular analyses use several benchmark datasets, while deep boundary stability is examined only on MNIST and CIFAR-10 with a finite set of independently trained teachers.
- **Operational scope.** Our segment-crossing measure is an approximate geometric proxy rather than an exact distance to the decision boundary, and our evaluation-composition experiments manipulate the composition of the evaluation set rather than claiming to capture arbitrary real-world distribution shifts.
- **Downstream scope.** Our counterfactual-transfer experiment is example-based and measures whether surrogate-selected changes transfer to the teacher; it should not be interpreted as establishing causal or actionable recourse.

6 Conclusion

Across several tabular datasets, surrogate families, and capacities, we find that teacher–surrogate disagreement is often concentrated near the teacher's decision boundary under both confidence-based and approximate segment-crossing definitions of boundary proximity. This structure is not merely descriptive: confidence-stratified fidelity profiles better predict fidelity when evaluation composition changes than the corresponding global score.

At the same time, boundary-aware fidelity is not a complete characterization of the teacher–surrogate relationship. Surrogates with similar global and boundary-local fidelity can exhibit sharply different example-based counterfactual transfer behavior, and independently trained deep teachers with high predictive agreement need not identify the same evaluation examples as belonging to their boundary regions. Therefore, fidelity should be interpreted as conditional on a teacher–surrogate pair and its evaluation space, rather than as a single definitive measure of behavioral equivalence.

Ultimately, for surrogate fidelity, we argue that alongside the question posed by Ribeiro et al. (2016), "*Why should I trust you?*", we must also ask: "*Where can I trust you?*"

AI Use Statement

In this work, we used generative AI tools to assist with literature discovery, method implementation and software development, and to support brainstorming, refinement of research framing, experimental design and evaluation, interpreting results, and drafting and editing the manuscript. We did not use AI to generate data or to formulate or prove mathematical claims. We manually reviewed relevant literature identified with AI assistance, reviewed and executed all AI-generated code, verified resulting analyses against our experiments, and critically evaluated AI-assisted research suggestions and interpretations. We take full responsibility for the final content of this work.

Reproducibility Statement

Experimental code is available at `https://anonymous.4open.science/r/boundary-aware-surrogate-fidelity-review-4D01`, and a deanonymized repository will be linked after acceptance.

References

Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1992.

Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1194–1206. ACM, 2022. doi: 10.1145/3531146.3533179.

Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296:103471, 2021. doi: 10.1016/j.artint.2021.103471.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.

Jackson Eshbaugh. "Faithful to What?" On the Limits of Fidelity-Based Explanations. In *The 2nd Workshop on Scientific Methods for Understanding Deep Learning (Sci4DL) at the Fourteenth International Conference on Learning Representations (ICLR)*, Rio de Janeiro, RJ, Brazil, 2026.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936.

Roel Henckaerts, Katrien Antonio, and Marie-Pier Côté. When stakes are high: Balancing accuracy and transparency with model-agnostic interpretable data-driven surrogates. *Expert Systems with Applications*, 202:117230, 2022. doi: 10.1016/j.eswa.2022.117230.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. In *Neural Information Processing Systems (NIPS) Deep Learning and Representation Learning Workshop*, 2015.

Van Quoc Phuong Huynh, Florian Beck, Josef Küng, and Johannes Fürnkranz. Globally valid rule-based explanations for black-box models. *Computational Intelligence*, 42(4):e70278, 2026. doi: 10.1111/coin.70278.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. In *3rd Workshop on Human Interpretability in Machine Learning (WHI 2018) at the 35th International Conference on Machine Learning (ICML 2018)*, Stockholm, Sweden, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Shiye Lei, Fengxiang He, Yancheng Yuan, and Dacheng Tao. Understanding deep learning via decision boundary. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):1533–1544, January 2025. doi: 10.1109/TNNLS.2023.3326654.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards Faithful Model Explanation in NLP: A Survey, January 2024.

Ettore Mariotti, Adarsa Sivaprasad, and Jose Maria Alonso Moral. Beyond prediction similarity: ShapGAP for evaluating faithful surrogate models in XAI. In *Explainable Artificial Intelligence*, volume 1901 of *Communications in Computer and Information Science*, pp. 160–173. Springer, 2023. doi: 10.1007/978-3-031-44064-9_10.

Rafael Poyiadzi, Xavier Renard, Thibault Laugel, Raul Santos-Rodriguez, and Marcin Detyniecki. Understanding surrogate explanations: The interplay between complexity, fidelity and coverage, July 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. In *The Third International Conference on Learning Representations (ICLR)*, San Diego, California, USA, 2015.

Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. GLocalX – from local to global explanations of black box AI models. *Artificial Intelligence*, 294:103457, 2021. doi: 10.1016/j.artint.2021.103457.

Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13689–13698, June 2022. doi: 10.1109/CVPR52688.2022.01333.

William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1993.

A Experimental Details

This appendix provides implementation details for the experiments reported in Sections 4.1–4.4. Unless otherwise stated, tabular experiments are repeated over ten random seeds. All feature-standardization statistics are estimated using the corresponding training partition only.

A.1 Datasets and Preprocessing

Our tabular experiments use four datasets distributed with `scikit-learn`, each interpreted as a binary classification task for the purposes of this study. Table 2 summarizes these tasks.

Dataset	Examples	Features	Binary task
Breast Cancer Wisconsin (Wolberg et al., 1993)	569	30	Original binary labels
Wine (Aeberhard & Forina, 1992)	178	13	Class 0 vs. remaining classes
Iris (Fisher, 1936)	150	4	Class 0 vs. remaining classes
Diabetes (Efron et al., 2004)	442	10	Target $\geqslant$ median vs. target $<$ median

Table 2: Tabular datasets used in the experiments and the binary classification tasks constructed from them. Wine and Iris are converted to one-vs-rest tasks using class 0 as the positive class. The Diabetes target is binarized at its median.

For each random seed, data partitions are generated using stratified sampling with respect to the binary ground-truth label. Features are standardized using a `StandardScaler` fit only on the training partition, and the resulting transformation is then applied unchanged to all held-out partitions. Experiment-specific partition sizes are described below.

A.2 Tabular Teacher Models

For all tabular experiments, the teacher is a binary multilayer perceptron with two hidden layers of widths 64 and 32, ReLU activations, and a single scalar output logit. Teachers are trained on the ground-truth binary labels using binary cross-entropy with logits and Adam with a learning rate

10^{-2}. Training is full-batch for 500 epochs. Predictions are obtained by thresholding the output logit at zero.

A fresh teacher is trained for each random seed. Before model construction and training, we seed Python, NumPy, and PyTorch random number generators, including CUDA when available. Teacher confidence in the binary experiments is measured using absolute logit magnitude, $c_T(x) = |f_T(x)|$.

A.3 Surrogate Models

All surrogates are trained on the training inputs using the corresponding teacher hard predictions as targets, rather than the ground-truth labels. Table 3 summarizes the representative surrogate configurations used in the cross-family analyses.

Surrogate family	Configuration
Decision tree	maximum depth 5
Logistic regression	$C = 1$, `liblinear` solver
k-NN	$k = 5$
MLP	one hidden layer, width 32, ReLU, `lbfgs`

Table 3: Representative surrogate configurations used in the cross-family robustness analyses.

Our primary boundary-fidelity capacity analysis varies decision-tree maximum depth over $\{1, 2, 3, 5, 7, 10\}$. The fidelity-prediction analysis uses depths $\{3, 5, 7\}$, while the counterfactual-transfer analysis uses $\{2, 3, 5, 7, 10\}$. Cross-family robustness analyses use the representative configurations in Table 3. Logistic regression and MLP surrogates are fit with a maximum of 5000 iterations; all other parameters use the corresponding `scikit-learn` defaults unless otherwise stated. Surrogate models with stochastic components use the experimental seed.

A.4 Boundary-Proximity Computation

For confidence-based proximity, we compute the absolute teacher logit $c_T(x) = |f_T(x)|$ for each evaluation example. For a requested boundary fraction α, the boundary region contains the $\lceil \alpha |X| \rceil$ examples with the smallest confidence values.

For the segment-crossing proxy, each evaluation example x is paired with the nearest observed evaluation example x' receiving the opposite teacher prediction, where distance is Euclidean in the standardized feature space. We then perform 30 iterations of binary search along the line segment joining x and x' to approximate a teacher decision crossing. Segment-crossing proximity is the Euclidean distance from x to this estimated crossing. This quantity is an operational geometric proxy and is not the exact minimum Euclidean distance from x to the teacher's full decision boundary.

If no oppositely predicted evaluation example exists, the segment-crossing distance is treated as undefined and the example is excluded from the valid geometric ranking. For segment-crossing proximity, the requested boundary fraction is applied to this valid ranking. Boundary regions are evaluated at $\alpha \in \{0.10, 0.20, 0.25, 0.40\}$. For a fixed teacher, random seed, and evaluation set, confidence-based and segment-crossing boundary regions are computed independently of the surrogate and reused across all surrogate configurations.

A.5 Deep Teacher Models

The deep-teacher stability analysis uses five independently trained teachers for each of MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009), corresponding to random seeds $\{0, 1, 2, 3, 4\}$.

For MNIST, we use a small convolutional network with three 3×3 convolutional layers with stride 1 and padding 1, containing 32, 64, and 128 channels. The first two convolutional layers are followed by 2×2 max-pooling. The convolutional representation is then mapped through a fully connected layer of width 256 with ReLU activation and dropout 0.5, followed by a 10-class output layer. Images are normalized using mean 0.1307 and standard deviation 0.3081. Models are trained for 30 epochs using cross-entropy loss and Adam with learning rate 10^{-3}.

For CIFAR-10, we use ResNet-18 models trained from scratch. To accommodate 32×32 inputs, the initial convolution is replaced with a 3×3 convolution with stride 1 and padding 1, and the initial max-pooling layer is removed. Training data are augmented using random 32×32 crops with four pixels of padding and random horizontal flips. Images are normalized using channel means $(0.4914, 0.4822, 0.4465)$ and standard deviations $(0.2023, 0.1994, 0.2010)$. Models are trained for 200 epochs using cross-entropy loss and SGD with learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, and cosine learning-rate annealing.

Both datasets use a batch size of 128. Python, NumPy, and PyTorch random number generators are seeded independently for each run; CUDA seeds are also set when available, and cuDNN deterministic execution is enabled. For the deep-teacher experiments, boundary proximity is operationalized using the margin between the largest and second-largest output logits, with smaller margins indicating greater proximity to the decision boundary.

B Additional Boundary-Fidelity Results

This section provides additional numerical detail for the boundary-local fidelity analysis in Section 4.1. We first report the decision-tree capacity sweep underlying the main boundary-local results, then extend the surrogate-family comparison across all evaluated boundary fractions.

B.1 Decision-Tree Capacity Sweep

Table 4 reports the numerical results underlying the decision-tree capacity analysis in Section 4.1. Values are averaged over ten random seeds. We report the 25% boundary region as a representative operating point; Figure 1 in the main text shows how the global-to-boundary gap varies across all evaluated boundary fractions.

Dataset	Depth	Global fidelity (%)	Confidence gap (pp)	Segment-crossing gap (pp)
Breast Cancer	1	91.6	12.3 ± 3.2	13.7 ± 2.5
	2	93.3	11.2 ± 3.0	10.8 ± 3.9
	3	93.4	11.8 ± 2.4	11.8 ± 3.0
	5	94.0	11.9 ± 2.2	10.1 ± 2.8
	7	94.1	12.9 ± 2.2	10.6 ± 2.4
	10	94.2	12.5 ± 2.0	10.9 ± 2.6
Diabetes	1	72.5	17.5 ± 5.8	16.0 ± 6.3
	2	75.0	18.6 ± 4.1	15.9 ± 4.4
	3	74.1	18.8 ± 4.7	13.8 ± 4.0
	5	73.6	18.6 ± 4.2	13.9 ± 4.6
	7	74.3	19.9 ± 3.4	15.2 ± 3.6
	10	72.4	19.8 ± 2.9	13.6 ± 3.9
Wine	1	90.6	7.0 ± 4.7	4.8 ± 5.7
	2	95.4	10.4 ± 5.5	8.9 ± 5.2
	3	95.0	10.0 ± 5.8	10.7 ± 4.9
	5	95.0	11.4 ± 5.2	11.4 ± 4.4
	7	95.0	11.4 ± 5.2	11.4 ± 4.4
	10	95.0	11.4 ± 5.2	11.4 ± 4.4
Iris	1	99.6	1.2 ± 1.6	1.2 ± 1.6
	2	99.6	1.2 ± 1.6	1.2 ± 1.6
	3	99.6	1.2 ± 1.6	1.2 ± 1.6
	5	99.6	1.2 ± 1.6	1.2 ± 1.6
	7	99.6	1.2 ± 1.6	1.2 ± 1.6
	10	99.6	1.2 ± 1.6	1.2 ± 1.6

Table 4: Decision-tree capacity sweep at boundary fraction $\alpha = 0.25$. Global fidelity and boundary gaps are averaged over ten random seeds. Boundary gaps are defined as $F_{\text{global}} - F_{\text{boundary}}$; gap values are shown as means with approximate 95% confidence intervals.

B.2 Surrogate-Family Robustness

Table 1 in the main text compares representative surrogate families at a boundary fraction of 25%. Figure 5 extends this analysis across all evaluated boundary fractions. The qualitative pattern is stable as the boundary fraction varies: Breast Cancer and Diabetes exhibit positive boundary-local gaps across all four surrogate families under both boundary definitions. On Wine, the effect is smaller for the non-tree surrogates, whose global fidelity is already near saturation, while Iris remains a near-null case throughout.

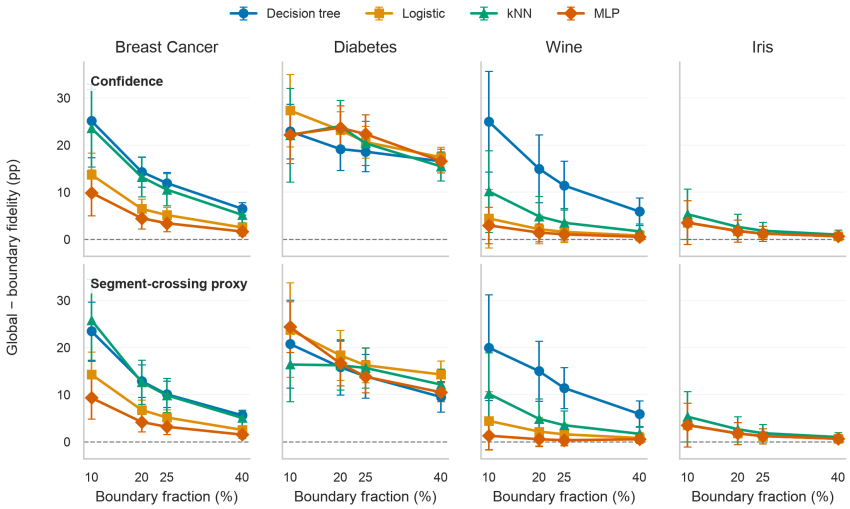

Figure 5: Surrogate-family robustness across boundary fractions. Global-to-boundary fidelity gaps are shown for representative decision-tree, logistic-regression, k-NN, and MLP surrogates under confidence-based (top) and approximate segment-crossing (bottom) measures of boundary proximity. Points show means over ten random seeds; error bars indicate approximate 95% confidence intervals.

C Additional Fidelity-Prediction Results

This section provides robustness analyses for the fidelity-prediction experiment in Section 4.2. We first examine sensitivity to the number of confidence bins used to construct the fidelity profile, then test whether the improvement over global fidelity persists across representative surrogate families.

C.1 Five-Bin Sensitivity

Our primary fidelity-prediction analysis constructs the conditional fidelity profile using three confidence bins. Table 5 compares this choice with a five-bin profile using the same source and target partitions. If a source confidence bin contains no examples, its conditional fidelity is set to the source-set global fidelity for purposes of constructing the profile prediction. Overall, the five-bin profile retains an advantage over the global-fidelity baseline, although its mean absolute error is slightly higher than that of the three-bin profile. The difference is concentrated on Wine, the smallest dataset; Breast Cancer is essentially unchanged, while Diabetes shows only a small increase in error. The corresponding overall reduction in mean absolute error relative to the global baseline is 39.7% with three bins and 34.6% with five bins.

Dataset	Global baseline (pp)	3-bin profile (pp)	5-bin profile (pp)
Breast Cancer	4.30	2.93	2.92
Diabetes	13.62	5.92	6.04
Wine	5.09	5.02	6.10
Overall	7.67	4.62	5.02

Table 5: Sensitivity of target-fidelity prediction to the number of confidence bins. Values are mean absolute prediction errors across surrogate depths, random seeds, and target compositions. The global-fidelity baseline is unchanged by the number of profile bins.

C.2 Surrogate-Family Robustness

We additionally repeat the three-bin fidelity-prediction analysis using representative decision-tree, logistic-regression, k-NN, and MLP surrogates. Table 6 reports results overall and separately by dataset. Averaged across datasets and target compositions, the confidence-stratified profile has lower mean absolute prediction error for every surrogate family. The improvement is especially pronounced on Diabetes. Breast Cancer also improves for three of the four families, while the MLP configuration is effectively unchanged; on Wine, several high-fidelity surrogates produce little or no difference between the two predictors.

Scope	Surrogate	Global baseline (pp)	3-bin profile (pp)	Improvement (pp)
Overall	Decision tree	8.00	5.06	2.95
	Logistic	6.78	2.96	3.82
	k-NN	7.26	4.07	3.19
	MLP	5.13	2.81	2.32
Breast Cancer	Decision tree	4.56	3.63	0.93
	Logistic	1.93	1.44	0.49
	k-NN	4.39	3.14	1.25
	MLP	1.23	1.29	-0.06
Diabetes	Decision tree	14.17	6.33	7.84
	Logistic	17.59	6.62	10.97
	k-NN	14.06	6.07	7.99
	MLP	13.33	6.30	7.03
Wine	Decision tree	5.28	5.21	0.07
	Logistic	0.83	0.83	0.00
	k-NN	3.33	2.99	0.35
	MLP	0.83	0.83	0.00

Table 6: Fidelity-prediction robustness across surrogate families. Values are mean absolute target-fidelity prediction errors. "Improvement" is computed before rounding as global baseline error minus profile error, so positive values favor the confidence-stratified profile.

D Additional Counterfactual-Transfer Results

This section provides additional detail for the example-based counterfactual-transfer analysis in Section 4.3. We first report results across all evaluated surrogate capacities, then examine specific surrogate pairs with closely matched observational fidelity but substantially different transfer behavior.

D.1 Full Configuration Results

Table 7 reports results for every evaluated decision-tree depth, averaged over ten random seeds. Global and boundary-conditioned fidelity can remain relatively stable across surrogate capacities even when counterfactual transfer rates differ. Across all configurations, the experiment contains

10,250 eligible transfer attempts; every eligible example has an available candidate with the opposite surrogate prediction.

Dataset	Depth	Global fidelity (%)	Boundary fidelity (%)	Transfer rate (%)
Breast Cancer	2	92.4 ± 2.1	79.7 ± 4.1	50.2 ± 12.5
	3	93.8 ± 1.7	83.1 ± 3.8	51.9 ± 11.1
	5	94.6 ± 1.5	84.8 ± 4.6	49.0 ± 14.2
	7	94.3 ± 1.7	84.5 ± 3.9	45.3 ± 13.5
	10	94.5 ± 1.7	84.5 ± 4.4	44.4 ± 13.9
Diabetes	2	73.9 ± 2.5	48.2 ± 5.5	50.2 ± 11.1
	3	73.4 ± 4.0	45.9 ± 4.9	47.4 ± 8.9
	5	71.4 ± 2.5	49.5 ± 5.9	46.1 ± 8.1
	7	71.2 ± 2.1	52.3 ± 8.3	46.4 ± 8.3
	10	71.7 ± 1.2	54.1 ± 7.4	48.2 ± 8.5
Wine	2	96.4 ± 1.8	88.9 ± 4.6	69.2 ± 15.3
	3	95.0 ± 1.6	82.2 ± 5.8	72.2 ± 14.3
	5	95.0 ± 1.6	82.2 ± 5.8	68.3 ± 13.0
	7	95.0 ± 1.6	82.2 ± 5.8	68.3 ± 13.0
	10	95.0 ± 1.6	82.2 ± 5.8	68.3 ± 13.0

Table 7: Counterfactual-transfer results across decision-tree capacities. Fidelity and transfer rates are reported as means across ten random seeds, with $\pm$ denoting approximate 95% confidence-interval half-widths. Boundary-local fidelity uses $\alpha = 0.25$.

D.2 Matched-Fidelity Examples

We define a matched pair as two surrogate configurations trained under the same dataset and random seed whose global fidelity and boundary-conditioned fidelity each differ by at most one percentage point. Among qualifying pairs, Table 8 reports the pair with the largest counterfactual-transfer gap for each dataset. For Breast Cancer, the depth-3/depth-7 pair is also the example highlighted in Figure 3.

Dataset	Seed	Depths	Global fidelity (%)	Boundary fidelity (%)	Transfer rates (%)	Gap (pp)
Breast Cancer	3	3 / 7	94.7 / 94.7	82.8 / 82.8	73.1 / 24.1	49.1
Diabetes	7	5 / 10	72.7 / 72.7	54.5 / 54.5	56.2 / 37.5	18.8
Wine	4	3 / 10	100.0 / 100.0	100.0 / 100.0	100.0 / 61.1	38.9

Table 8: Matched-fidelity surrogate pairs with differing counterfactual-transfer behavior. Each row compares two decision-tree surrogates trained under the same dataset and random seed. The transfer gap is the difference between the two transfer rates in percentage points.

These examples make the limitation of observational fidelity particularly clear. In each case, the matched surrogates have the same global and boundary-conditioned fidelity at the reported precision, yet changes selected using the surrogate transfer to the teacher at substantially different rates. The Wine example is especially stark: both surrogates achieve perfect global and boundary-conditioned fidelity, while their transfer rates differ by nearly 39 percentage points. Thus, even highly localized observational agreement does not uniquely determine behavior under surrogate-selected changes.

E Additional Deep-Teacher Stability Results

This section provides additional numerical detail for the deep-teacher stability analysis in Section 4.4. We first report the predictive accuracies of the independently trained teachers, then provide the pairwise stability statistics underlying Figure 4.

E.1 Teacher Accuracies

Table 9 reports test accuracy for each independently trained teacher. The teachers within each dataset achieve closely matched overall predictive performance, providing a comparable basis for the pairwise boundary-stability analysis.

Dataset	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4
MNIST	99.09	99.27	99.33	99.18	99.39
CIFAR-10	95.30	95.27	95.35	95.42	95.63

Table 9: Test accuracy (%) of independently trained deep teachers. Each teacher is evaluated on the common 10,000-example test set for its dataset.

E.2 Pairwise Boundary-Set Stability

For each dataset, we compare all ten unique pairs among the five independently trained teachers. Table 10 reports the mean pairwise statistics at each evaluated boundary fraction α. Global agreement is computed over the full test set. Boundary-set overlap is measured by Jaccard similarity, while union and intersection agreement measure predictive agreement on the union and intersection, respectively, of the two teachers' boundary sets.

Dataset	α	Global agr. (%)	Jaccard (%)	Union agr. (%)	Intersection agr. (%)	Global − union (pp)
MNIST	0.10	99.18	39.39	94.31	86.63	4.87
	0.20	99.18	43.13	97.08	93.32	2.10
	0.25	99.18	45.01	97.63	94.77	1.55
CIFAR-10	0.10	96.05	44.32	72.09	58.39	23.96
	0.20	96.05	56.81	84.50	74.63	11.54
	0.25	96.05	58.53	87.47	79.50	8.58

Table 10: Pairwise stability of independently trained deep teachers. Values are means across the ten unique teacher pairs for each dataset.

The numerical results reinforce the distinction between global predictive agreement and stability of the examples identified as lying near the decision boundary. At the 10% boundary fraction, pairwise global agreement is 99.18% on MNIST and 96.05% on CIFAR-10, while mean boundary-set Jaccard overlap is only 39.39% and 44.32%, respectively. Predictive agreement also falls within these regions, particularly on CIFAR-10, where agreement on the union of the two teachers' boundary sets is 72.09%. As the boundary fraction increases, overlap and boundary-region agreement increase, but independently trained teachers continue to identify substantially different sets of examples as lying near the decision boundary despite their high global predictive agreement.