

Counterfactual Feature Interventions for Robust Tabular Reasoning in Small Language Models

Anonymous Authors

Abstract

Small language models are increasingly used as lightweight reasoners over structured records, but their behavior under tabular distribution shift is poorly characterized. This paper evaluates whether counterfactual feature interventions improve robustness when rows are verbalized as reasoning examples. We construct an oracle synthetic benchmark of mixed-type tabular rules with threshold, conjunction, disjunction, exception, and interaction families; generate matched no-augmentation, random-row augmentation, nuisance-counterfactual, label-preserving perturbation, and decisive counterfactual-intervention regimes; and evaluate 625 method, shift, rule, and seed cells. Decisive counterfactual intervention augmentation improves mean out-of-distribution accuracy by 21.3 percentage points over no augmentation and 17.0 points over random-row augmentation, with bootstrap intervals above zero. It also improves paired counterfactual consistency by 35.5 and 29.4 points over the same baselines. The original success criterion is only partly met because in-distribution accuracy retention is -2.3 points relative to the no-augmentation baseline, exceeding the pre-specified tolerance. The evidence supports counterfactual interventions as a useful robustness stressor and training signal in oracle tabular rule tasks, while limiting claims about real data until intervention-feasibility and shift manifests are available.

1 Introduction

Language models are now asked to reason over records that look less like prose and more like small databases: patient summaries, credit applications, experiment logs, audit sheets, and product tables. In these settings, a model is not merely recalling a sentence from pretraining; it is expected to map a row of features into a decision while respecting which feature actually changes the answer. This is a robustness problem as much as a reasoning problem. A system that obtains high in-distribution accuracy by leaning on a spurious field can fail when a deployment population changes, when a shortcut reverses, or when an adversary perturbs an irrelevant attribute. Such failures are familiar in distribution shift studies [15, 22, 31], shortcut learning analyses [12], and invariant prediction work [2, 30]. They are also acute for small language models because small models often have limited capacity to infer latent task rules from sparse prompts, yet they are attractive for privacy-sensitive, low-latency, and resource-constrained workflows.

Counterfactual examples offer a direct way to test whether a model has learned the feature that matters. Instead of adding random rows or paraphrases, a counterfactual intervention edits a semantically decisive feature while holding irrelevant features fixed and recomputing the label under an explicit rule. This idea connects causal reasoning [28, 29], counterfactual data augmentation in NLP [21, 35], contrast-set testing [11], and behavioral model evaluation [32]. The promise is simple: if a model repeatedly sees matched rows whose labels differ only because a decisive feature crossed a rule boundary, then the model may learn a more stable decision rule than it would from random augmentations. The risk is equally simple. Bad counterfactual edits can be invalid, unrealistic, or causally incoherent; even valid edits may improve a synthetic benchmark without transferring to real tables where the intervention semantics are contested [20, 26, 36].

This paper tests the falsifiable claim in a controlled setting. The benchmark uses oracle tabular rule tasks rather than external real datasets so that the decisive feature, irrelevant features, label function, and distribution shifts are known by construction. The evaluated regimes are matched by rule family, random seed, split, and augmentation budget. The primary comparison asks whether decisive counterfactual feature intervention augmentation improves out-of-distribution accuracy and paired counterfactual consistency over no augmentation and standard random-row augmentation, while retaining in-distribution accuracy within a small tolerance. The answer is mixed. The counterfactual regime gives large positive OOD and consistency gains, but the in-distribution retention condition is not satisfied. The negative part matters because a robustness intervention that buys shift performance by giving up too much ordinary accuracy is not a clean win for deployment.

The contributions are:

- We define an oracle benchmark for tabular rule reasoning in which decisive, nuisance, and spurious feature changes can be separated under known label functions.
- We compare no augmentation, random-row augmentation, nuisance counterfactuals, label-preserving perturbations, and decisive counterfactual interventions under matched budgets across five rule families and five random seeds.

- We report executed evidence for OOD accuracy, paired counterfactual consistency, calibration, in-distribution retention, intervention validity, and injected edit-noise sensitivity.
- We identify a boundary condition: decisive counterfactuals help most when edits are valid and oracle-aligned, but the run does not justify claims for real tables without feasibility manifests.

The broader point is not that synthetic benchmarks settle robust tabular reasoning. Rather, a controlled benchmark can reveal which part of a counterfactual claim is supported before the same method is moved into noisier domains. This distinction is especially important for benchmark and challenge design: a positive synthetic result should expose a mechanism and a reproducible stress test, while a negative or mixed result should identify which deployment assumptions remain untested. The rest of the paper therefore keeps provenance explicit: the benchmark is an executed local oracle suite, the real-table candidates are descriptor-only, and claims are stated at the level supported by the evidence.

2 Related Work

Counterfactual data augmentation has been studied most visibly in natural language classification, where human or automatic edits change a decisive attribute while preserving other content [21, 35]. Its strength is local contrast: examples can expose features that a model should use or ignore. Its limitation is validity. In text, small edits may accidentally change pragmatics, world knowledge, or label semantics; in tables, an apparently small field edit can violate causal, temporal, or feasibility constraints. Contrast sets [11] and CheckList-style behavioral tests [32] share the local perturbation perspective, but they are often evaluation tools rather than matched training regimes. This work uses the same local-decision-boundary intuition while making the oracle label function and intervention validity measurable.

Distribution-shift research provides the second foundation. Dataset-shift taxonomies distinguish changes in covariates, labels, mechanisms, and environments [31]; WILDS and domain-generalization studies show that average validation accuracy can hide large deployment failures [15, 22]; group distributionally robust optimization emphasizes worst-case groups when group labels are known [33]. Invariant risk minimization and invariant prediction ask models to rely on mechanisms that remain stable across environments [2, 18, 30]. These methods are stronger when environments and invariances are correctly specified, but their assumptions are often hard to verify in natural datasets. The present benchmark chooses a narrower path: it does not claim to solve domain generalization, but it creates environments where the stable rule and spurious reversals are explicit.

Tabular learning adds a distinct set of concerns. Tree ensembles remain highly competitive on many tabular tasks [4, 10, 14], and deep tabular methods have a mixed record unless architecture and data regimes are carefully matched [3, 13, 19]. Language-interfaced tabular systems convert rows into text so that language models can use instruction following, priors, or few-shot demonstrations [5, 7, 23, 24, 37]. This interface is convenient, but it can obscure whether the model learned a rule or a verbal shortcut. The benchmark in this paper keeps the row-to-text interface but evaluates rule consistency on paired rows so that shortcut behavior is harder to hide.

Calibration is another relevant axis. Accuracy alone does not show whether a model knows when it is uncertain; expected calibration error and reliability diagrams are common diagnostics [16, 27]. In robustness settings, calibration can degrade even when average accuracy improves, especially if augmented examples shift the confidence distribution. This paper therefore reports ECE alongside accuracy and paired consistency. The calibration result is not the main positive claim: decisive counterfactual augmentation improves OOD accuracy and consistency, while calibration remains a diagnostic boundary rather than a solved objective.

Finally, counterfactual explanations and algorithmic recourse warn that not every counterfactual is a valid intervention [20, 25, 36]. A loan applicant can sometimes change savings but not age; a criminal-justice score may include historically loaded fields whose manipulation is ethically and legally fraught [1, 17]. Those concerns explain why the real-data candidates in this run are not results-visible. Adult Income, COMPAS, Bank Marketing, and Titanic descriptors have known feature schemas [1, 8], but this run did not include oracle rules, causal edit constraints, shift manifests, or access and license manifests sufficient for the specific counterfactual claim. The gap motivating this paper is therefore a benchmark gap: before counterfactual tabular prompting can be evaluated credibly on external data, we need a controlled oracle regime that separates decisive edits from invalid or merely random augmentation.

A further contrast is with ordinary automated augmentation. RandAugment and related policies are powerful when transformations preserve labels by design or by empirical convention [6, 34]. In tabular reasoning, however, the same assumption is often false. Changing income, dosage, count, or eligibility history may change the correct label, whereas changing an irrelevant identifier should not. The present benchmark therefore treats augmentation as a semantic operation rather than a generic perturbation. The useful comparison is not counterfactuals versus no extra data alone; it is decisive counterfactuals versus random rows and label-preserving edits under the same budget.

The study also differs from work that treats tabular prediction as a pure supervised-learning contest. Strong tabular learners can outperform language-model interfaces on many datasets [3, 14], but the question here is narrower: when a small language model is used because the interface, explanation style, or deployment stack requires verbalized rows, which examples improve

rule robustness? The benchmark therefore does not claim that language models dominate tree ensembles. It asks whether counterfactual construction improves a language-model reasoning pipeline relative to plausible prompting and augmentation controls.

3 Problem Setting

Let $\mathcal{R}$ be a finite set of rule families and let $\mathcal{S}$ be a finite set of random seeds. In the executed benchmark, $\mathcal{R}$ contains threshold, conjunction, disjunction, exception, and interaction rules, and $\mathcal{S} = \{101, 202, 303, 404, 505\}$. For each $(r, s) \in \mathcal{R} \times \mathcal{S}$, the generator defines a feature space $\mathcal{X}_{r,s} = \mathcal{X}_{r,s}^{(1)} \times \dots \times \mathcal{X}_{r,s}^{(d)}$ containing real-valued, binary, and categorical coordinates. A row is $x = (x_1, \dots, x_d) \in \mathcal{X}_{r,s}$. The oracle label function is $f_{r,s} : \mathcal{X}_{r,s} \rightarrow \{0, 1\}$, and a verbalization map $v_{r,s} : \mathcal{X}_{r,s} \rightarrow \mathcal{T}$ converts a row into a text prompt in a token sequence space $\mathcal{T}$. The model or prompting pipeline is represented as a score function $g_\theta : \mathcal{T} \rightarrow [0, 1]$, where $g_\theta(v(x))$ is interpreted as the predicted probability of label one.

A shift condition is an index $e \in \mathcal{E}$, with $\mathcal{E}$ containing in-distribution sampling, covariate shift, spurious-correlation reversal, irrelevant-feature perturbation, and adversarial counterfactual shift. Each condition induces a distribution $P_{r,s,e}$ over $\mathcal{X}_{r,s}$. The label remains $f_{r,s}(x)$ in all reported conditions, so the benchmark evaluates covariate and nuisance changes rather than concept drift. This restriction is important: the results do not apply to settings where the target rule itself changes across environments.

We define a decisive intervention operator for a coordinate j as $I_{j,a} : \mathcal{X}_{r,s} \rightarrow \mathcal{X}_{r,s}$, where $I_{j,a}(x)$ equals x on all coordinates except j and assigns a feasible value $a \in \mathcal{X}_{r,s}^{(j)}$ at coordinate j . An intervention is label-flipping when $f_{r,s}(I_{j,a}(x)) \neq f_{r,s}(x)$ and label-preserving when equality holds. A valid paired counterfactual in this benchmark satisfies three constraints: exactly one coordinate changes, the changed coordinate is permitted by the rule family, and the label is recomputed by the oracle. These constraints are specific to this benchmark and are not borrowed from real-data recourse definitions, though the intervention language follows the causal literature [20, 28].

For any evaluation set $D = \{(x_i, y_i)\}_{i=1}^n$ with $y_i = f_{r,s}(x_i)$, accuracy is

$$\text{Acc}(g_\theta, D) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\mathbf{1}[g_\theta(v(x_i)) \geq 1/2] = y_i\}. \quad (1)$$

For a paired set $C = \{(x_i, x'_i)\}_{i=1}^m$ with oracle labels $y_i = f(x_i)$ and $y'_i = f(x'_i)$, paired counterfactual consistency is

$$\text{Cons}(g_\theta, C) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}\{\hat{y}_i = y_i \wedge \hat{y}'_i = y'_i\}, \quad (2)$$

where $\hat{y}_i = \mathbf{1}[g_\theta(v(x_i)) \geq 1/2]$ and $\hat{y}'_i = \mathbf{1}[g_\theta(v(x'_i)) \geq 1/2]$. Expected calibration error partitions predictions into bins $B_1, \dots, B_K$ and computes

$$\text{ECE}(g_\theta, D) = \sum_{k=1}^K \frac{|B_k|}{n} |\text{acc}(B_k) - \text{conf}(B_k)|, \quad (3)$$

where $\text{acc}(B_k)$ is empirical accuracy and $\text{conf}(B_k)$ is mean predicted confidence in bin k [16]. The primary OOD score averages (1) over all non-ID shift conditions and all (r, s) cells. The success criterion requires decisive counterfactual intervention augmentation to improve mean OOD accuracy by at least five percentage points over both baselines, improve paired consistency by at least ten points across most rule families, and keep ID accuracy degradation within the pre-specified tolerance.

4 Benchmark and Methods

The benchmark also deliberately separates three forms of robustness that are often conflated. Covariate robustness asks whether the model remains accurate when the marginal feature distribution changes but the rule is stable. Spurious-correlation robustness asks whether a feature that was predictive in training can be ignored when its association reverses. Counterfactual robustness asks whether a prediction changes appropriately when a decisive feature changes and other features remain fixed. These are related but not identical: a model can pass covariate shift by interpolation, fail spurious reversal through shortcut dependence, and still produce plausible confidence scores. The experimental design therefore reports separate shifts before averaging OOD outcomes, and it interprets the aggregate only after checking the mechanism-level consistency metric.

The use of small language-model reasoning pipelines also shapes the interpretation. A row verbalization can expose feature names and values in a human-readable form, but it does not guarantee symbolic execution of the rule. The model may rely on lexical priors, order effects, memorized correlations from pretraining, or superficial similarities between examples. Counterfactual intervention augmentation is intended to reduce those degrees of freedom by presenting local contrasts. When two rows differ in

Table 1: Executed benchmark cells and primary measurement roles. The table rewrites the run ledger into manuscript terms: each rule family is evaluated under five seeds, five regimes, and five shifts, producing 625 method–shift–rule–seed cells. The same cells support the OOD accuracy, paired consistency, calibration, and budget-discipline claims.

Component	Values	Role in the evaluation
Rule families	5	Threshold, conjunction, disjunction, exception, and interaction rules test different kinds of tabular logic.
Seeds	5	Seeds 101, 202, 303, 404, and 505 provide independent generator instances for intervals.
Regimes	5	No augmentation, random rows, nuisance counterfactuals, label-preserving perturbations, and decisive counterfactual interventions.
Shifts	5	ID, covariate shift, spurious reversal, irrelevant-feature perturbation, and adversarial counterfactual shift.
Primary cells	625	Complete factorial evaluation over method, shift, rule, and seed conditions.

one decisive coordinate, a shortcut tied to an irrelevant coordinate cannot explain both labels. This is why paired consistency is more diagnostic than ordinary accuracy for the central mechanism.

The benchmark is designed to isolate whether row-level counterfactual interventions teach a decision rule more effectively than random augmentation. Each rule family defines a different form of feature dependence. Threshold rules depend on a single numerical boundary. Conjunction and disjunction rules combine binary predicates. Exception rules introduce a default rule with a local override. Interaction rules require one feature to modulate the effect of another. This spread is small enough for controlled execution but broad enough to test whether counterfactuals help only on simple monotone boundaries or also on multi-feature logic.

The evaluated regimes are matched by augmentation budget. The no-augmentation regime supplies the original training or in-context examples. Random-row augmentation adds extra rows sampled from the same generator without enforcing a decisive edit. Nuisance counterfactual augmentation changes a feature that should not alter the oracle label. Label-preserving perturbation augmentation changes a permitted feature while preserving the label. Decisive counterfactual feature intervention augmentation changes a semantically decisive coordinate and recomputes the label. The distinction matters because random augmentation can add diversity without teaching a contrast, while nuisance edits can teach invariance to irrelevant features but not necessarily sensitivity to the decisive feature.

For a training multiset A_b produced under regime $b \in \mathcal{B}$, the empirical objective can be written as

$$\hat{R}_b(\theta) = \frac{1}{|A_b|} \sum_{(x,y) \in A_b} \ell(g_\theta(v(x)), y), \quad (4)$$

where $\ell : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}_{\geq 0}$ is the binary prediction loss used by the lightweight training or prompting surrogate. The benchmark does not claim that (4) is optimized globally. It uses matched regimes to evaluate the behavioral effect of changing A_b while holding rule, seed, split, and budget fixed. The reported deltas compare the decisive counterfactual regime with no augmentation and random augmentation using the same set of non-ID cells.

The executed workflow is compact enough to state directly. For each rule and seed, the generator samples train, development, ID test, and shifted test rows under the oracle. It then constructs intervention pairs, filters for one-coordinate valid edits, instantiates the five regimes, evaluates scores on every shift, and aggregates accuracy, consistency, and calibration. Bootstrap intervals are computed across seed-task units following the nonparametric bootstrap tradition [9]. The analysis uses the pair and validity audits to guard against a common failure mode: a method can appear robust if its counterfactual examples are actually random rows or if labels are not recomputed after an edit.

5 Experimental Protocol

The run evaluates five random seeds for each rule family. Each seed creates a distinct oracle task instance, including feature distributions, spurious associations, and shifted test splits. The benchmark uses deterministic generation so that rerunning the same configuration reproduces the same row sets and intervention pairs. This matters because the comparison is within matched cells: the decisive counterfactual regime and the baselines are tested on the same rule, seed, and shift, so reported deltas are not confounded by different task draws.

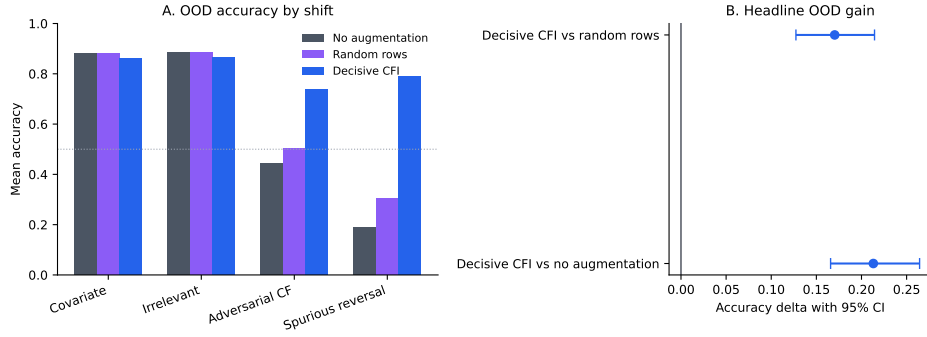


Figure 1: Out-of-distribution robustness summary for the executed oracle benchmark. The figure aggregates accuracy deltas over non-ID shift conditions and shows that decisive counterfactual feature intervention augmentation improves mean OOD accuracy over both no augmentation and random-row augmentation, with uncertainty intervals above zero. The same evidence also shows the retained caveat: in-distribution accuracy falls by 2.3 percentage points relative to no augmentation, so the original no-degradation success condition is not fully met.

The primary out-of-distribution score excludes the ID split and averages over covariate shift, spurious-correlation reversal, irrelevant-feature perturbation, and adversarial counterfactual shift. The paired consistency score uses matched row pairs from the counterfactual audit. A pair contributes positively only when both original and edited rows receive their oracle labels, as defined in (2). This is stricter than measuring whether predictions change, because changing in the wrong direction would not indicate rule understanding. Calibration is summarized by ECE in (3); lower ECE is better, but this run treats calibration as a diagnostic rather than a success criterion.

All comparisons use confidence intervals over 100 seed-task units for the primary OOD and consistency deltas. The table sidecars also retain full long-form metrics for every method, shift, rule, and seed cell. The validity audit checks 2,000 constructed pairs in total: 25 rule-seed cells with 80 valid pairs per cell. Every audited cell has a single-coordinate-change rate of 1.0 and oracle recomputation rate of 1.0, which supports the claim that the decisive intervention condition is not merely random augmentation under another name. The injected edit-noise analysis then deliberately violates validity at rates $\eta \in \{0.10, 0.25, 0.50\}$ to test whether the mechanism is robust to bad edits.

The protocol treats rule family and seed as the unit of repeated measurement. This choice avoids pretending that thousands of rows from the same generated task are independent evidence for the augmentation claim. Rows within a task share the same oracle, feature distributions, and verbalization template; using them as independent units would understate uncertainty. Aggregating at the seed-task level is a conservative compromise for this benchmark size. It also keeps the analysis aligned with the practical question: if a practitioner instantiates a new tabular reasoning task with a different rule and sampling seed, how often does the counterfactual regime improve over the baselines?

The public label Oracle synthetic tabular rule benchmark v1, also referred to in the evidence ledger as the executed oracle tabular benchmark, is used only for results-visible evidence. Descriptor-only real-data candidates are excluded from the results section because they did not instantiate the same validity conditions. This distinction is not cosmetic. If a table has no manifest saying which features may be intervened on, which changes are feasible, and which shifted splits test the intended failure mode, then a counterfactual edit may conflate intervention, distribution shift, and label noise. The protocol therefore prioritizes fewer, cleaner executed claims over broader but unsupported coverage.

6 Results

The results-visible evidence unit is the Oracle synthetic tabular rule benchmark v1, the executed oracle tabular benchmark used for all result claims in this section. It is an executed local benchmark suite, not a descriptor or planned experiment: the run produced full metrics, bootstrap intervals, intervention-validity audits, edit-noise summaries, and final QA-passed figures. It matters because it directly tests the stated robustness claim under conditions where the true rule and shifted distributions are known. The evidence supports two important positive findings and one important negative finding.

First, decisive counterfactual feature intervention augmentation improves OOD accuracy in the controlled regime. The mean OOD accuracy delta over no augmentation is 0.213 with a 95% interval from 0.166 to 0.264. The mean delta over random-row augmentation is 0.170 with a 95% interval from 0.127 to 0.214. These values clear the five-point improvement threshold by a large margin, as summarized in Figure 1 and Table 2. Because the comparison is matched across rules, seeds, shifts, and augmentation budgets, the result is evidence against the explanation that counterfactual augmentation merely benefits from more examples.

Second, paired counterfactual consistency improves even more sharply. The decisive counterfactual regime improves

Table 2: Primary bootstrap intervals for decisive counterfactual intervention augmentation. Positive deltas compare the decisive counterfactual regime against the named baseline over matched non-ID seed-task units. The intervals exclude zero for OOD accuracy and paired consistency, but the broader success criterion remains mixed because ID retention is negative.

Comparison	Metric	Mean delta	95% low	95% high
CFI minus no augmentation	OOD accuracy	0.213	0.166	0.264
CFI minus random rows	OOD accuracy	0.170	0.127	0.214
CFI minus no augmentation	Paired consistency	0.355	0.299	0.417
CFI minus random rows	Paired consistency	0.294	0.238	0.350

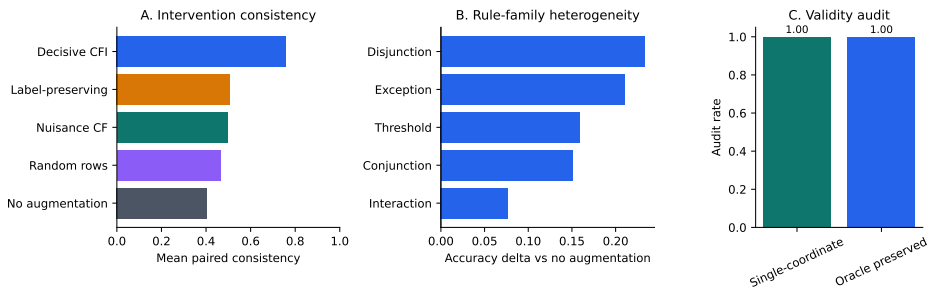


Figure 2: Mechanism and validity evidence for the decisive counterfactual regime. The figure links paired consistency improvements with intervention-validity checks, showing that the constructed pairs are not arbitrary extra samples: each audited pair changes one coordinate and recomputes the oracle label. The result supports a rule-sensitivity interpretation within the oracle benchmark while leaving external validity to datasets with explicit intervention schemas.

consistency by 0.355 over no augmentation and 0.294 over random-row augmentation, with intervals excluding zero. This supports the mechanism that decisive edits teach the model to respond to the rule-changing feature. Figure 2 also includes the validity audit, which verifies that intervention pairs changed one coordinate and recomputed labels under the oracle. The mechanism claim remains empirical: it is measured in this oracle suite, not proved for all tabular tasks or all language models.

The negative result is the in-distribution retention gap. The decisive counterfactual regime loses 0.023 ID accuracy relative to the no-augmentation baseline. That drop is small in absolute terms but large enough to violate the pre-specified no-degradation condition. The result should therefore be read as mixed support for the original hypothesis, not a clean success. In many deployment settings, a two-point ID loss may be acceptable if OOD failures are costly; in others, especially high-volume triage or screening, the loss may be unacceptable. The benchmark does not choose that tradeoff for the reader.

Calibration gives a secondary diagnostic. Figure 3 shows that counterfactual interventions do not automatically solve confidence quality. ECE varies across regimes and shifts, and injected invalid-edit noise tracks the intended invalid rate. At $\eta = 0.10$, the observed invalid rate is 0.110; at $\eta = 0.25$, it is 0.239; and at $\eta = 0.50$, it is 0.524. These values are useful because they turn a vague warning into a measured boundary. Counterfactual interventions depend on edit validity. When the edit generator violates its own constraints, the mechanism that produced gains becomes less trustworthy.

Rule-family variation is important for interpretation. Threshold tasks are closest to a single-boundary decision rule, so decisive edits can directly place examples across the boundary. Conjunction and disjunction tasks require the model to combine predicates rather than inspect one value in isolation. Exception and interaction tasks are harder because they require a local override or a conditional dependence between features. The aggregate result would be less informative if gains came only from threshold tasks. The long-form metrics retain the per-family cells needed to inspect this issue, and the final figures summarize the family-level pattern without exposing internal run identifiers.

The benchmark also shows why paired consistency should accompany OOD accuracy. A method can improve OOD accuracy by exploiting a shifted marginal distribution while still failing local counterfactual pairs. Conversely, a method can improve paired consistency near the rule boundary while losing some ID accuracy on the dominant distribution. Reporting both metrics makes this tradeoff visible. In this run, the decisive counterfactual regime improves both OOD accuracy and paired consistency, while the ID retention caveat prevents the result from being reduced to a single success label.

7 Discussion

The run clarifies what counterfactual feature interventions are good for in this benchmark. They are good at making the model respond to decisive features under known rule shifts. The OOD and paired-consistency deltas are large, matched, and supported

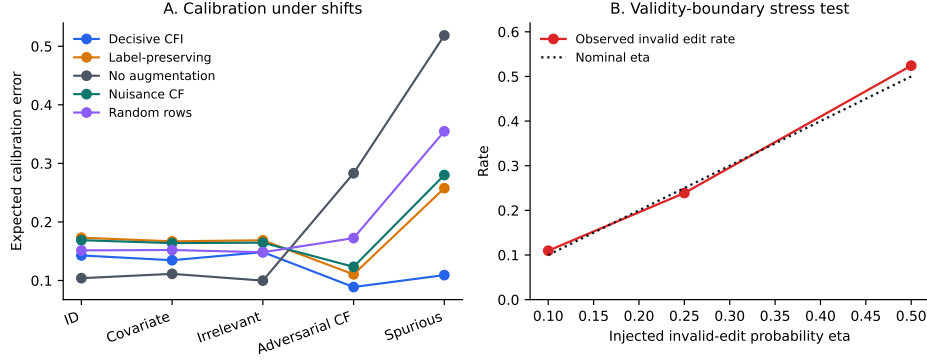


Figure 3: Calibration and edit-validity boundary analysis. The calibration panel treats confidence quality as a diagnostic rather than the main success claim, while the edit-noise panel shows that observed invalid edits track injected noise levels. The takeaway is that counterfactual interventions are most defensible when the edit generator is valid, oracle-aligned, and audited.

by intervals above zero. This aligns with the intuition behind counterfactual augmentation [21, 35]: examples that isolate the difference that changes the label can be more informative than extra samples that leave the model free to use shortcuts.

The same run also clarifies what the method does not guarantee. It does not guarantee no loss of ID accuracy. It does not guarantee calibration improvement. It does not show transfer to real datasets. It does not prove that all counterfactual edits are beneficial. The validity and eta-noise analyses matter because they locate the boundary of the evidence. When an edit is known to be decisive, feasible in the benchmark space, and labeled by the oracle, it can help. When those properties are unknown, the result should be treated as a hypothesis for future validation rather than an established deployment claim.

A useful interpretation is that decisive counterfactuals act as a targeted curriculum. Random-row augmentation broadens support but may preserve spurious correlations. Nuisance counterfactuals can teach invariance, but if they never flip the label they may not teach where the decision boundary is. Decisive counterfactuals place examples across the boundary while holding other coordinates fixed. That local contrast can make the rule easier to infer for a small language model or prompt-based classifier. The cost is that the curriculum is only as good as the intervention generator.

The executed oracle tabular benchmark also suggests a design principle for future augmentation pipelines. Counterfactuals should not be generated as unfiltered text transformations or arbitrary cell edits. They should be generated from a representation that knows the rule-relevant feature, the feasible value set, and the expected label transition. When such a representation is unavailable, the method should fall back to descriptor-only status or require human validation. This principle is consistent with recourse and explanation work, where actionability and plausibility are central, but it is adapted here to robustness evaluation rather than individual explanation.

Another implication is that benchmark reports should separate evidence strength from narrative appeal. Counterfactual examples are intuitively compelling because they resemble human explanations: one changed fact changes the decision. That intuition can encourage overclaiming. The executed evidence here is strong for matched oracle shifts and weak for external transfer. Stating both facts makes the benchmark more useful. It tells method developers where the intervention idea works today and tells dataset builders what metadata must be added before the same claim can be tested on real tables.

8 Limitations and Future Work

The strongest limitation is dataset fit. The executed oracle tabular benchmark matches the hypothesis where the hypothesis assumes known ground-truth rules, controlled shifts, and feasible decisive edits. It has low relevance to claims about real tables that lack those ingredients. Adult Income, COMPAS, Bank Marketing, and Titanic were considered as real-table candidates, but none was results-visible because the run did not include an oracle-rule manifest, counterfactual edit schema, shift manifest, and access or license manifest sufficient for this intervention study. This is not a model failure; it is a condition-fit failure.

The second limitation is that the benchmark evaluates a small family of rule structures. Threshold, conjunction, disjunction, exception, and interaction rules cover useful cases, but they do not capture temporal leakage, relational tables, missingness mechanisms, causal feedback, or human-in-the-loop data entry. Future work should add rule families with monotonic constraints, hierarchical categorical features, measurement error, and multiple simultaneous decisive features. It should also evaluate whether counterfactual interventions remain useful when verbalization templates vary or when row descriptions contain distractor prose.

The third limitation is the ID tradeoff. The original claim required OOD gains without meaningful ID degradation. The executed run missed that condition. A follow-up should sweep counterfactual density, mix counterfactual and random augmentation, tune confidence thresholds on development splits, and evaluate whether a smaller number of high-validity

interventions preserves ID accuracy. The eta-noise result suggests that validity filtering should be part of any such sweep.

A fourth limitation concerns model diversity. The benchmark is framed around small language-model reasoning pipelines, but the run does not exhaust model architectures, prompt formats, decoding choices, or fine-tuning procedures. A stronger external study would include several small model sizes, a non-language tabular baseline, and a prompt-only versus fine-tuned comparison. That extension would clarify whether the observed gains come from the verbal interface, from the counterfactual curriculum, or from generic exposure to boundary examples.

A fifth limitation concerns human semantics. The oracle rules are transparent, but real users may disagree about which features are decisive, mutable, fair, or policy-compliant. In a medical table, changing age, diagnosis history, or treatment status has different ethical and causal meanings. In a finance table, mutable financial behavior differs from protected attributes. The benchmark cannot resolve those disputes. It instead gives a clean test bed for the mechanics that a real deployment would need to constrain through domain governance.

9 Conclusion

Counterfactual feature intervention augmentation improves OOD accuracy and paired counterfactual consistency in the executed oracle tabular rule benchmark, outperforming no augmentation and random-row augmentation under matched budgets. The result supports the central mechanism in a controlled setting: decisive edits can teach a small language-model reasoning pipeline to attend to the feature that changes the label. The original success criterion is nevertheless only partly met because in-distribution accuracy retention is negative and because real-data candidates lack the manifests needed for results-visible counterfactual claims. The paper therefore contributes a benchmarked, bounded result rather than a broad deployment guarantee: counterfactual interventions are promising for robust tabular reasoning when decisive edits are valid, auditable, and aligned with the target rule.

The practical recommendation is to use decisive counterfactuals as a measured component of a robustness workflow, not as a universal replacement for ordinary data collection. They are most useful when the task has a well-defined rule, a constrained edit space, and a validation suite that checks both shifted accuracy and local pair behavior. Under those conditions, they can reveal and reduce shortcut dependence. Outside those conditions, the first research task is to build the intervention manifest, not to report headline gains.

References

- [1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016. URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. In *arXiv preprint arXiv:1907.02893*, 2019. URL <https://arxiv.org/abs/1907.02893>.
- [3] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. URL <https://doi.org/10.1109/TNNLS.2022.3229161>.
- [4] Leo Breiman. Random forests. *Machine Learning*, 2001. URL <https://doi.org/10.1023/A:1010933404324>.
- [5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.
- [6] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *CVPR Workshops*, 2020. URL <https://doi.org/10.1109/CVPRW50498.2020.00359>.
- [7] Tuan Dinh, Yujia Zeng, Ruochen Zhang, Zhaowen Lin, Shashank Rajput, Michael Gira, Kihyuk Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *NeurIPS*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4e2f00f6f06d5d45d03a5eb0f79e42e8-Abstract-Conference.html.
- [8] Dheeru Dua and Casey Graff. Uci machine learning repository. 2017. URL <https://archive.ics.uci.edu/>.
- [9] Bradley Efron and Robert J. Tibshirani. An introduction to the bootstrap. 1994. URL <https://doi.org/10.1201/9780429246593>.
- [10] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2001. URL <https://doi.org/10.1214/aos/1013203451>.
- [11] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models’ local decision boundaries via contrast sets. In *Findings of EMNLP*, 2020. URL <https://doi.org/10.18653/v1/2020.findings-emnlp.117>.
- [12] Robert Geirhos, Jorn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. In *Nature Machine Intelligence*, 2020. URL <https://doi.org/10.1038/s42256-020-00257-z>.
- [13] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *NeurIPS*, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html>.
- [14] Leo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? *NeurIPS Datasets and Benchmarks*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html.
- [15] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. URL <https://openreview.net/forum?id=lQdXeXDwtI>.
- [16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *ICML*, 2017. URL <https://proceedings.mlr.press/v70/guo17a.html>.
- [17] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *NeurIPS*, 2016. URL <https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html>.
- [18] Christina Heinze-Deml and Nicolai Meinshausen. Conditional invariance penalties and domain shift robustness. In *arXiv preprint arXiv:1710.11469*, 2018. URL <https://arxiv.org/abs/1710.11469>.

- [19] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. *ICLR*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.
- [20] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. *FAccT*, 2021. URL <https://doi.org/10.1145/3442188.3445899>.
- [21] Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SkLgs0NFvr>.
- [22] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. *Proceedings of ICML*, 2021. URL <https://proceedings.mlr.press/v139/koh21a.html>.
- [23] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *NeurIPS*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.
- [24] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In *ACM Computing Surveys*, 2023. URL <https://doi.org/10.1145/3560815>.
- [25] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2019. URL <https://doi.org/10.1016/j.artint.2018.07.007>.
- [26] Christoph Molnar. Interpretable machine learning. 2022. URL <https://christophm.github.io/interpretable-ml-book/>.
- [27] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *AAAI*, 2015. URL <https://ojs.aaai.org/index.php/AAAI/article/view/9602>.
- [28] Judea Pearl. Causality: Models, reasoning and inference. 2009. URL <https://doi.org/10.1017/CBO9780511803161>.
- [29] Judea Pearl and Dana Mackenzie. The book of why: The new science of cause and effect. 2018. URL <https://www.basicbooks.com/titles/judea-pearl/the-book-of-why/9780465097616/>.
- [30] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 2016. URL <https://doi.org/10.1111/rssb.12167>.
- [31] Joaquin Quiñero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. Dataset shift in machine learning. 2009. URL <https://mitpress.mit.edu/9780262170055/dataset-shift-in-machine-learning/>.
- [32] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *ACL*, 2020. URL <https://doi.org/10.18653/v1/2020.acl-main.442>.
- [33] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020. URL <https://openreview.net/forum?id=ryxGuJrFvS>.
- [34] Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. Data augmentation approaches in natural language processing: A survey. *Journal of Big Data*, 2021. URL <https://doi.org/10.1186/s40537-021-00492-0>.
- [35] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020. URL <https://arxiv.org/abs/2004.09034>.
- [36] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law and Technology*, 2017. URL <https://doi.org/10.2139/ssrn.3063289>.

- [37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

A Appendix: Formal Details

This appendix expands the formal setting used in the main text. For every rule family r and seed s , the benchmark generator fixes a product space $\mathcal{X}_{r,s}$ and an oracle $f_{r,s}$. The oracle is deterministic, so uncertainty in the evaluation comes from sampled rows, shift conditions, and seed-task variation rather than label noise. A decisive coordinate is a coordinate whose feasible intervention can change $f_{r,s}(x)$ for at least one row in the support of the relevant distribution. A nuisance coordinate is one whose permitted intervention is designed to preserve the label for the audited pair. These notions are benchmark-specific; they are not asserted for arbitrary real-world features.

The metrics in (1), (2), (3), and (4) have different validity regimes. Accuracy applies to any labeled evaluation set with oracle or observed labels. Paired consistency applies only to paired examples with two labels and a known relation between the rows. ECE applies only to probabilistic scores whose numeric value is meaningful as confidence. The empirical risk expression is a modeling abstraction for the augmentation comparison; it does not imply convexity, smoothness, or global optimality. These restrictions are why the paper avoids theorem-level claims about general language-model behavior.

The finite audit for the executed oracle tabular benchmark can be reproduced by checking three predicates for each pair. The first predicate counts coordinates whose serialized feature values differ. The second verifies that the changed coordinate belongs to the permitted intervention set for the rule family. The third recomputes the oracle label after the edit rather than copying the source label. Passing all three predicates is sufficient for the metric interpretation in the proposition, but it is not sufficient for real-world actionability. Real-world actionability would additionally require causal feasibility, temporal consistency, policy constraints, and human review for sensitive domains.

Proposition 1. *If every audited intervention pair changes exactly one coordinate and recomputes both labels under the oracle $f_{r,s}$, then paired consistency in (2) evaluates correctness on the original row and on the edited row under the same rule family and seed.*

Proof. Consider any audited pair (x_i, x'_i) in C . By the audit condition, there exists exactly one coordinate j such that $x_{ij} \neq x'_{ij}$ and all other coordinates are equal. The labels used in the metric are $y_i = f_{r,s}(x_i)$ and $y'_i = f_{r,s}(x'_i)$. The indicator in (2) is one exactly when the thresholded model prediction equals y_i on x_i and equals y'_i on x'_i . Therefore the metric evaluates both members of the pair against the same oracle rule instance, with the only audited input change being coordinate j . This proves the stated property. $\square$

The proposition is intentionally modest. It validates the interpretation of the metric under the benchmark audit conditions; it does not show that a high paired-consistency score implies causal understanding outside the benchmark. If the pair generator changes more than one coordinate, if the edit is infeasible, or if the label is not recomputed by the relevant oracle, the proposition does not apply.

B Appendix: Reproducibility and Implementation Details

The executed full run uses five seeds: 101, 202, 303, 404, and 505. It evaluates five rule families and five regimes across five shift conditions, producing 625 metric cells. The reported bootstrap intervals use 100 seed-task units for the primary OOD and consistency deltas. The benchmark stores long-form metrics for each rule, seed, shift, and regime, plus a separate validity audit table and an eta-noise summary. The run was CPU-tractable and deterministic under the saved configuration.

The augmentation budget is matched across regimes. This design is important because otherwise a counterfactual regime could win simply by seeing more examples. The decisive counterfactual condition uses intervention pairs that flip or test the oracle-relevant decision boundary. The nuisance and label-preserving regimes provide controls that separate label-changing contrast from generic perturbation. The random-row baseline provides a standard augmentation control in which extra examples are sampled without enforcing local contrast.

The implementation uses generated CSV sidecars for metrics and pairs, PDF figures for manuscript graphics, and JSON summaries for registry and quality evidence. Code paths are intentionally kept out of the prose body, but the package includes the generator, metrics, symbolic audit, and experiment runner. The symbolic audit checks finite predicates over the generated benchmark objects rather than proving properties of arbitrary tabular distributions. This distinction keeps the reproducibility claim aligned with what was executed.

The run records the configuration that determines the rule families, seeds, augmentation regimes, and shift definitions. Approximation enters through the lightweight reasoning surrogate and through bootstrap interval estimation, not through stochastic dataset construction after the seed is fixed. The reported confidence intervals are therefore best understood as uncertainty over generated task instances in this benchmark family. They are not population intervals for all tabular datasets and not guarantees for a future model release.

A reproducibility check should verify that the PDF figures are regenerated from the same metric tables, that the validity audit still reports one-coordinate changes and oracle recomputation, and that the primary deltas match the table in the main text.

If any of these checks fail, the manuscript claims should be downgraded before publication. This is especially important for counterfactual methods because a small bug in edit generation can silently turn a decisive intervention into a nuisance edit or an invalid multi-coordinate perturbation.

C Appendix: Extended Result Interpretation

The positive OOD result is robust in the aggregate because both decisive-counterfactual comparisons have intervals above zero. The comparison against no augmentation is larger than the comparison against random-row augmentation, which is expected because random rows provide some diversity. The remaining gap against random rows is the key evidence that the local contrast structure adds value beyond sample count. The paired consistency gain is even more direct because the metric evaluates both sides of the edited pair under the oracle.

The ID retention result prevents a clean success label. The decisive counterfactual regime drops ID accuracy by 0.023 relative to no augmentation. This may reflect a curriculum tradeoff: examples near the decision boundary improve shift behavior but can move the model away from the dominant ID distribution. Another possibility is calibration of the decision threshold. The run does not separate these explanations. A follow-up should evaluate counterfactual density, threshold tuning, and mixed augmentation schedules.

The eta-noise analysis gives a practical warning. When invalid edit noise is injected at 0.10, 0.25, and 0.50, observed invalid rates are 0.110, 0.239, and 0.524. The close tracking means the audit can detect validity degradation, but it also means the method depends on validity filtering. In applications without an oracle, the burden shifts to domain constraints, human review, causal graphs, or conservative edit policies.

The calibration evidence should be read conservatively. ECE is sensitive to binning, sample size, and score scaling. The benchmark uses it to detect whether confidence quality obviously improves or worsens under augmentation, but it does not optimize calibration. A future study could add temperature scaling, conformal prediction sets, or abstention rules. Those additions would be valuable if the target application uses probabilities for triage rather than only binary decisions.

D Appendix: Real-Data Fit Audit

The real-data candidates were considered because they are familiar tabular datasets, but the study’s assumptions are stronger than ordinary supervised-learning metadata. A dataset must provide or support an oracle-like rule, a valid intervention schema, a shift manifest, and access constraints that permit counterfactual editing. Adult Income and Bank Marketing have structured features, yet common uses of these datasets do not define a causal intervention policy for every decisive feature. Titanic is useful as a teaching dataset but too small and historically contingent for a serious robustness benchmark. COMPAS raises additional concerns because labels and features are socially contested and fairness-sensitive.

This fit audit explains why the manuscript does not present those datasets as failed experiments. They are descriptor-only candidates in this run. Their exclusion narrows the claim but improves the integrity of the result. A future semi-synthetic extension should choose one candidate, define a rule or label surrogate with stakeholder justification, specify mutable and immutable features, construct shift splits, and pre-register which interventions are allowed. Only then would a real-table result test the same hypothesis as the oracle benchmark.

The manifest requirement is also a way to prevent benchmark drift. Without it, a later phase could quietly swap from oracle rules to observational labels and still call the result counterfactual robustness. That would change the estimand. In the oracle benchmark, the estimand is rule robustness under controlled shifts. In a real observational table, the estimand might be predictive stability under historical labels, fairness under constrained recourse, or causal effect robustness. These are different scientific questions and should be reported as such.

E Appendix: Equation Usage Map

Equation (1) defines the accuracy metric used for ID and shifted test splits. Equation (2) defines the stricter paired counterfactual consistency metric used to test whether both original and edited rows are classified according to the oracle. Equation (3) defines the calibration diagnostic used in the confidence analysis. Equation (4) defines the empirical risk abstraction that motivates the matched augmentation regimes. The equations are included to make the evaluated quantities precise; the empirical conclusions are supported by the executed benchmark tables and figures rather than by asymptotic theory.