A SHORTCUT CAN BE REAL YET UNDIAGNOSABLE: IDENTIFIABILITY LIMITS OF OBSERVATIONAL SHORTCUT AUDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

A shortcut can be real yet observationally undiagnosable. Shortcut auditing is therefore an identification problem before it is an estimation problem: a fixed predictor may rely on identity-associated information even when the data lack the same-state cross-identity comparisons needed to reveal that reliance. We define predictor-specific reliance under an explicit target comparison distribution and distinguish full-target reliance, the support-restricted population estimand, and the finite audit estimate. If the target assigns positive mass to unsupported comparisons, full-target reliance need not be point-identified. Under bounded prediction discrepancy, partial support yields an identified interval whose width contracts with target-weighted support coverage. This result motivates a support-aware audit that reports prediction contrast, admissible pair count, and target coverage separately, and returns ABSTAIN when no admissible requested comparison is available. In a controlled intervention that varies only auditor-visible support while holding the predictor and finite full-target oracle reference fixed, recovery fidelity improves in $39/40$ model–seed endpoints, while all $800/800$ zero-support views abstain. In CWRU vibration sensing, support thinning degrades recovery of the full-support observational reference in all five repetitions. At a fixed pair count, concentrated support incurs $12.4\times$ the recovery error of broadly distributed support at the median seed. Stanford battery EIS provides a complementary boundary case: substantial deployment degradation does not identify its source. Shortcut audits must therefore establish which predictor comparisons are observationally supported before interpreting estimated reliance.

1 INTRODUCTION

A model can rely on a shortcut that an observational audit cannot diagnose Geirhos et al. (2020); Ilyas et al. (2019); Sagawa et al. (2020); McCoy et al. (2019). This problem is especially relevant in physical sensing, where measurements often encode both task state and acquisition-specific factors such as device, batch, route, protocol, or source Zech et al. (2018); Badgeley et al. (2019); DeGrave et al. (2021). These factors may be highly decodable without affecting a trained predictor, while deployment degradation may coincide with their shift without identifying them as the mechanism of failure Koh et al. (2021); Sagawa et al. (2019). Shortcut auditing must therefore distinguish identity presence, predictor-specific reliance, and deployment harm Hewitt & Liang (2019); Ravichander et al. (2021); Elazar et al. (2021); Belinkov (2022). The central question is whether the observations contain the comparisons needed to determine how a fixed predictor responds to identity-associated variation.

We formulate shortcut auditing as an identification problem. We define full-target predictor-specific reliance under an explicit distribution over same-state cross-identity comparisons and distinguish it from the support-restricted population estimand and its finite-sample estimate. If a relevant alternative identity is unsupported at a target state, the same observational law and fixed predictor can be compatible with different values of full-target reliance. Under partial support and bounded prediction discrepancy, we derive an identified interval whose width contracts with target-weighted support coverage. This result connects classical overlap and partial identification to fixed-predictor behavior rather than population treatment effects D'Amour et al. (2021); Manski (1990). It also

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

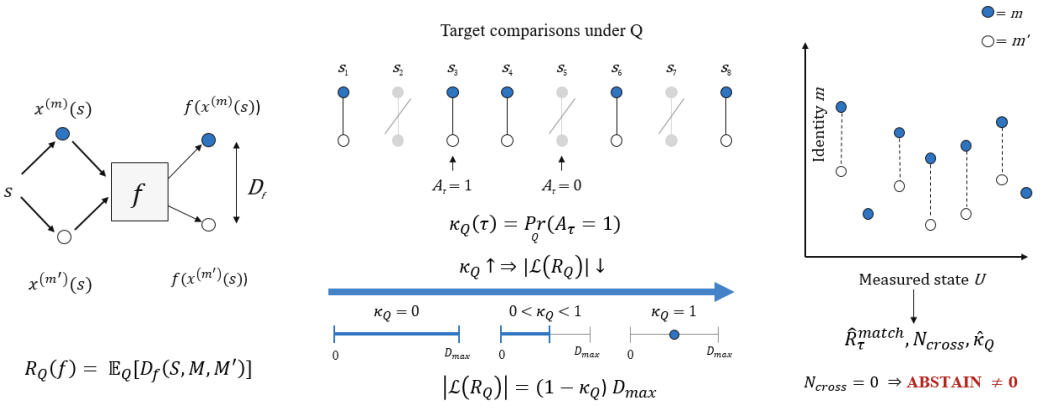


Figure 1: **Identification structure of observational shortcut auditing.** (a) Full-target reliance compares a fixed predictor across identities at the same task state. (b) Target-weighted support coverage controls the width of the identified set. (c) A finite audit reports supported prediction contrast, admissible pair count, and target coverage, and returns ABSTAIN when no admissible comparison exists.

motivates a support-aware audit that reports prediction contrast, admissible comparison count, and target coverage separately, and returns ABSTAIN when the required comparison is unsupported. Figure 1 summarizes the resulting target–support–identification–audit chain.

We evaluate this framework in controlled and real sensing settings. A controlled intervention varies only auditor-visible cross-identity support while holding the trained predictor, oracle reliance, deployment degradation, audit-set size, and task-state and identity marginals fixed. Increasing support improves oracle-recovery fidelity in $39/40$ model–seed endpoints, while all $800/800$ zero-support views return ABSTAIN. State matching also substantially improves recovery of the oracle-reliance ordering, although the experiments do not establish that matching is necessary for coarse harmful-versus-harmless discrimination. Stanford battery EIS provides an attribution boundary: substantial deployment degradation does not by itself identify its mechanism. In CWRU vibration sensing, reducing state-region support monotonically degrades recovery of the full-support observational reference in all five training repetitions. At the same cross-identity pair count, concentrated support incurs $12.4\times$ the recovery error of broadly distributed support at the median seed, with the same direction in all five repetitions. These real-data analyses remain observational and do not identify causal shortcut effects.

Our contributions are:

- **Support-dependent identification.** We distinguish full-target reliance, support-restricted reliance, and finite-sample estimation, and derive a coverage-dependent identified interval under bounded prediction discrepancy.
- **Support-aware observational auditing.** We report prediction contrast, admissible pair count, and target coverage separately, and return ABSTAIN when the required cross-identity comparison is unsupported.
- **Controlled and real-data evidence.** Controlled interventions show that observational support governs recovery of a fixed reliance target, while Stanford and CWRU expose complementary boundaries in attribution, observational recovery, and support coverage.

2 Related Work

Shortcut evidence and predictor-specific reliance. Shortcut learning studies predictors that exploit predictive but brittle signals under distribution shift (Geirhos et al., 2020; Hermann et al., 2024). Existing audits use attribute predictability, controlled evaluation, dataset structure, or generalization

behavior to identify suspected shortcuts (Brown et al., 2023; Drenkow et al., 2026). Such evidence can show that an attribute is available, predictive, or associated with failure, but it does not generally determine how a particular frozen predictor responds to that attribute at comparable task states. We study this predictor-specific quantity and ask when observational data identify it.

From encoded information to functional use. Interpretability methods distinguish information encoded in a representation from information that affects model behavior. Amnesic probing removes encoded information (Elazar et al., 2021), while causal and counterfactual approaches intervene on linguistic or conceptual attributes to estimate behavioral effects (Amini et al., 2023; Feder et al., 2021; Abraham et al., 2022; Vigneshwaran et al., 2025). These methods construct the comparison of interest through intervention, transformation, or counterfactual generation. We study the observational setting in which the required comparison may instead be absent from support. The resulting question precedes estimation: whether the available observations identify the predictor contrast at all.

Observational support and partial identification. Overlap and positivity characterize whether observational comparisons are supported within relevant covariate strata, while partial-identification methods characterize what remains learnable when support is incomplete (Petersen et al., 2012; D'Amour et al., 2021; Manski, 1990). Related work also shows that observational evidence may remain compatible with distinct shortcut mechanisms and that explicit interventions or structural counterfactuals can resolve otherwise ambiguous explanations (Wang, 2026; Toker et al., 2026). We specialize this support problem to the behavior of a fixed predictor. Missing same-state identity support can leave full-target reliance non-identifiable, while partial support exposes only a support-restricted population quantity. Our audit makes this distinction operational by reporting predictor contrast, comparison count, and target-weighted support coverage separately, and by abstaining when the required comparison is unavailable.

3 When Is Shortcut Reliance Diagnosable?

An observational shortcut audit must distinguish three objects: the predictor behavior of interest, the part supported by the observational law, and the statistic computed from finite data.

3.1 Target Reliance and Observational Support

Let S denote the task-relevant physical state, M an identity- or acquisition-associated variable, and X the observed measurement:

$$X \sim p(x \mid S, M), \qquad Y = g(S), \qquad \hat{Y} = f(X). \tag{1}$$

The auditor observes a state description U, which need not fully determine S. For a fixed predictor f, let $x^{(m)}(s)$ denote an identity-specific realization at state s. We define the same-state cross-identity discrepancy as

$$D_f(s, m, m') = d_{\mathcal{Y}}\Big(f(x^{(m)}(s)), f(x^{(m')}(s))\Big), \qquad m \neq m'. \tag{2}$$

For a target distribution Q over requested comparisons (s, m, m'), the full-target predictor-specific reliance is

$$R_Q(f) = \mathbb{E}_Q\left[D_f(S, M, M')\right]. \tag{3}$$

This target is defined independently of the audit dataset. It differs from identity decodability, which concerns whether M can be recovered from X, and from deployment error, which measures predictive risk under shift.

The observational data need not support every comparison weighted by Q. Let $A_\tau(s, m, m') \in \{0, 1\}$ indicate whether a requested comparison is supported at state resolution τ, and define its target-weighted support coverage as

$$\kappa_Q(\tau) = \Pr_Q(A_\tau = 1). \tag{4}$$

When $\kappa_Q(\tau) > 0$, the support-restricted population estimand is

$$R_Q^{\sup}(f; \tau) = \mathbb{E}_Q\left[D_f(S, M, M') \mid A_\tau = 1\right]. \tag{5}$$

In general, $R_Q^{\sup}(f; \tau) \neq R_Q(f)$.

3.2 Support Determines the Identified Set

We first consider the favorable case $U = S$ and hold the predictor f fixed.

Proposition 3.1 (Support-dependent identification of predictor-specific reliance)**.** *Suppose $U = S$, f is fixed and nonconstant, and no structural assumption determines predictor responses on unsupported same-state identity branches.*

(i) *If Q assigns positive probability to an unsupported branch, the observational law need not point-identify $R_Q(f)$. In particular, there exist admissible worlds W_0 and W_1 with the same observational law and fixed predictor but different full-target reliance:*

$$P_{\text{obs}}^{W_0}(X, S, M) = P_{\text{obs}}^{W_1}(X, S, M), \qquad R_Q^{W_0}(f) \neq R_Q^{W_1}(f). \tag{6}$$

(ii) *If $0 \leq D_f \leq D_{\max}$ and $0 < \kappa_Q(\tau) \leq 1$, then*

$$R_Q(f) = \kappa_Q R_Q^{\sup}(f; \tau) + (1 - \kappa_Q)\mathbb{E}_Q[D_f \mid A_\tau = 0]\,. \tag{7}$$

Without further restrictions on unsupported comparisons, this decomposition yields the identified interval

$$R_Q(f) \in \left[\kappa_Q R_Q^{\sup}(f; \tau),\ \kappa_Q R_Q^{\sup}(f; \tau) + (1 - \kappa_Q)D_{\max}\right]. \tag{8}$$

The proposition yields three regimes. When $\kappa_Q = 1$, the identified interval collapses and $R_Q^{\sup} = R_Q$. When $0 < \kappa_Q < 1$, the observations identify $R_Q^{\sup}$ but leave R_Q partially identified. When $\kappa_Q = 0$, $R_Q^{\sup}$ is undefined and boundedness alone gives $R_Q(f) \in [0, D_{\max}]$. The identified-set width is $(1-\kappa_Q)D_{\max}$. Additional physical, parametric, generative, or invariance assumptions may tighten this set, but the resulting identification would depend on those assumptions rather than on P_{obs} alone. Appendix A gives the proof and a constructive non-identification witness.

3.3 A Support-Aware Finite Audit

Given observations (x_i, u_i, m_i), we construct deterministic one-to-one cross-identity matches at state resolution τ:

$$\mathcal{M}_\tau^{\text{cross}} = \{(i, j) : m_i \neq m_j,\ d_u(u_i, u_j) \leq \tau\}_{\text{1-to-1}}\,, \qquad N_{\text{cross}} = |\mathcal{M}_\tau^{\text{cross}}|\,. \tag{9}$$

For $N_{\text{cross}} > 0$, the state-matched estimator is

$$\widehat{R}_\tau^{\text{match}}(f) = \frac{1}{N_{\text{cross}}} \sum_{(i,j) \in \mathcal{M}_\tau^{\text{cross}}} d_{\mathcal{Y}}(f(x_i), f(x_j))\,. \tag{10}$$

This estimator summarizes predictor variation only over the comparisons available to the audit. We therefore report

$$\left(\widehat{R}_\tau^{\text{match}}(f),\ N_{\text{cross}},\ \widehat{\kappa}_Q(\tau)\right), \tag{11}$$

where $\widehat{\kappa}_Q$ is the empirical target-weighted support coverage. The three quantities measure prediction contrast, admissible pair count, and coverage of the target comparison space, respectively. Section 4 specifies $\widehat{\kappa}_Q$ for each evaluation setting.

When $N_{\text{cross}} = 0$, the audit returns ABSTAIN rather than zero. A zero estimate means that supported comparisons show no prediction difference, whereas ABSTAIN means that the required comparison is unavailable. Positive N_{cross} also does not imply full-target identification because the observed pairs may cover only a restricted part of Q. Matching on U improves state comparability but does not make the contrast causal because U need not exhaust the latent physical state S. Appendix B.1 gives the matching and support-qualification details.

4 Evaluation Design

We evaluate four links in the identification chain from Section 3. A controlled study varies observational support while holding predictor-specific reliance fixed. An estimator ablation tests whether state matching improves recovery of that target. Stanford EIS examines deployment degradation when attribution remains unidentified. CWRU evaluates support-qualified predictor sensitivity under full observational support, controlled support thinning, and fixed-pair-count coverage controls.

4.1 CHANGE SUPPORT, HOLD RELIANCE FIXED

Our first experiment operationalizes the support dimension of Proposition 3.1: does changing only observational support alter recovery of a fixed target? A controlled generator provides the complete requested set of same-state cross-identity comparisons and thereby defines a finite full-target oracle reference, $\widehat{R}_Q^{\mathrm{oracle}}(f)$. We then restrict which state anchors retain observable cross-identity support while keeping the predictor and oracle reference fixed.

We use 10 independent seeds, four predictor families—Ridge, Random Forest, MLP, and a small 1D CNN—and 20 state anchors per seed. For $H \in \{0, 2, 4, 6, 10, 16, 20\}$ supported anchors, we generate 20 repeated observational masks and define the design support fraction as $q = H/20$. We keep q distinct from the audit-reported target-weighted coverage $\widehat{\kappa}_Q$. The deterministic construction yields $N_{\mathrm{cross}} = 6H$.

Across support levels, the trained predictor, $\widehat{R}_Q^{\mathrm{oracle}}(f)$, deployment degradation, audit-set size, task-state marginal, and identity marginal remain fixed. Only the observable comparison structure changes. We measure recovery by the error of $\widehat{R}_\tau^{\mathrm{match}}(f)$ relative to $\widehat{R}_Q^{\mathrm{oracle}}(f)$ and measure stability across repeated masks. Seeds are independent replication units, whereas masks are repeated observational views. When $N_{\mathrm{cross}} = 0$, the audit returns ABSTAIN rather than zero. Appendix B.2 provides the controlled support construction and replication details.

4.2 TESTING MATCHING AND ESTIMATOR STRUCTURE

We next test whether state matching improves recovery when admissible cross-identity comparisons exist. We compare four estimators: an unmatched contrast, a pair-count-matched unmatched contrast, the state-matched estimator from Section 3.3, and state matching with same-identity subtraction. This comparison separates the effect of state comparability from raw comparison count and from the subtraction term.

The primary endpoint is Spearman correlation with $\widehat{R}_Q^{\mathrm{oracle}}$, which measures recovery of the continuous reliance ordering across benchmark conditions. AUROC and AUPRC are secondary endpoints for coarse harmful-versus-harmless discrimination and are analyzed separately from the ordering objective. Auxiliary synthetic benchmark worlds provide the evaluation conditions for the estimator ablation and test distinctions among identity observability, predictor reliance, and deployment degradation. Appendix B.3 summarizes these conditions and the corresponding evaluation pools.

4.3 STANFORD EIS: DEPLOYMENT STRESS TEST

Stanford EIS tests whether deployment degradation alone identifies its underlying mechanism. The dataset contains 848 usable spectra from 24 cells with state-of-health labels. We evaluate random, cell-held-out, temporal-held-out, and route-held-out protocols using the frozen regression pipeline. These splits change multiple aspects of the data distribution and do not isolate identity at fixed physical state. Stanford therefore quantifies deployment degradation but does not identify predictor-specific shortcut reliance.

4.4 CWRU: REAL-DATA SUPPORT AND DEPLOYMENT EVALUATION

CWRU evaluates the same support-aware logic under a different sensing mechanism and prediction task. We use five independent seeds and 12 fixed model–target-domain conditions per seed. Seeds are replication units, whereas conditions are repeated within-seed observations. We fix the state definition, identity definition, target distribution Q, predictor outputs, and matching procedure before evaluation. Total variation is the primary discrepancy metric, and Jensen–Shannon divergence is a predefined robustness metric.

Full-support observational reference. When all requested observational state regions are available, the matched estimator defines a *full-support observational reference* for predictor sensitivity. This quantity is neither oracle reliance nor a causal effect. We compare its ordering across model–domain conditions with independently measured deployment degradation. This comparison tests deployment relevance without treating degradation as attribution evidence.

Table 1: **Controlled support intervention.** Increasing observational support from sparse to full improves both recovery fidelity and stability of the fixed oracle reference.

Outcome	Improved endpoints	Mean reduction	95% CI	Median trajectory ρ
Recovery error	$39/40$	0.00325	$[0.00292, 0.00355]$	-0.829
Across-mask variability	$40/40$	0.00525	$[0.00494, 0.00559]$	-0.943

Support thinning. We then vary only the state-region support visible to the auditor, using $\widehat{\kappa}_Q \in \{0, 0.25, 0.5, 0.75, 1\}$. Predictor parameters, predictions, deployment outcomes, state and identity definitions, Q, and the full-support reference remain fixed. We measure error relative to the full-support reference, recovery of its cross-condition ordering, and variability across supported-region placements.

Coverage versus pair count. Finally, we fix $N_{\text{cross}} = 300$ and compare broad target-space coverage with the same number of matches concentrated in fewer state regions. This experiment directly tests $N_{\text{cross}} \neq \widehat{\kappa}_Q$: equal pair counts need not cover the comparisons requested by Q equally well. Appendix C provides the real-data preprocessing, reconstruction, support-thinning, and statistical details.

5 Results

We first test whether observational support controls recovery of a fixed predictor-specific reliance. We then examine estimator structure and target-space coverage before turning to deployment attribution and support-aware recovery on real sensing data.

5.1 More Support Improves Recovery of a Fixed Reliance

More observational support yields more faithful recovery of the fixed oracle reference (Figure 2). From $q = 0.1$ to $q = 1.0$, absolute recovery error decreases in $39/40$ model–seed endpoint comparisons, with a mean reduction of 0.00325 (95% CI $[0.00292, 0.00355]$). All $40/40$ nonzero-support trajectories exhibit a negative association between support and recovery error, with median within-trajectory Spearman $\rho = -0.829$. The predictor and $\widehat{R}_Q^{\text{oracle}}(f)$ remain fixed across these views, so the observed differences concern recoverability rather than the underlying reliance.

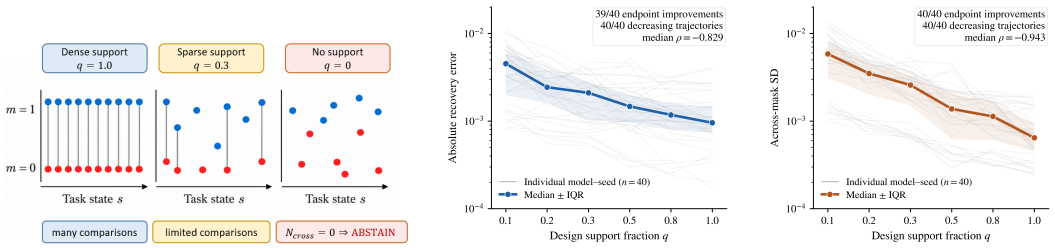
Support also improves recovery stability. Across-mask variability decreases from sparse to full support in all $40/40$ model–seed endpoint comparisons, with a mean reduction of 0.00525 (95% CI $[0.00494, 0.00559]$). All $40/40$ support–variability trajectories are negatively associated with support, with median Spearman $\rho = -0.943$. These results indicate that finite-audit stability depends on support placement, with placement sensitivity decreasing as support increases. Table 1 summarizes the endpoint and trajectory-level results.

Zero support differs qualitatively from low support. All $800/800$ views with $q = 0$ contain no admissible same-state cross-identity comparison and therefore return Abstain. This outcome is not a zero reliance estimate: a zero estimate requires an observed comparison with no prediction difference, whereas Abstain indicates that the requested comparison is unavailable.

5.2 State Matching Improves Oracle Reliance Recovery

State matching substantially improves recovery of the continuous oracle-reliance ordering (Table 2). The state-matched estimator reaches Spearman $\rho = 0.988$, compared with 0.312 for the unmatched estimator and 0.283 for the count-matched unmatched baseline. This improvement is not reproduced by matching comparison count alone. Adding same-identity subtraction yields $\rho = 0.880$, providing no demonstrated additional benefit for continuous oracle-order recovery.

The coarse harmful-versus-harmless task shows a mixed pattern. The unmatched estimator attains higher AUROC than the state-matched estimator (0.921 versus 0.840), whereas the state-matched estimator attains slightly higher AUPRC (0.810 versus 0.796). These results do not establish that



(a) **(b)** **(c)**

Figure 2: **Observational support controls recovery of a fixed predictor-specific reliance.** **(a)** We vary the design support fraction q while keeping the predictor and finite full-target oracle reference fixed. Dense support provides many admissible same-state cross-identity comparisons, sparse support provides fewer comparisons, and zero support yields $N_{\text{cross}} = 0$ and ABSTAIN. **(b)** Absolute recovery error decreases with support in all $40/40$ nonzero-support trajectories, with $39/40$ endpoint improvements and median within-trajectory Spearman $\rho = -0.829$. **(c)** Across-mask variability also decreases with support in all $40/40$ trajectories and endpoint comparisons, with median Spearman $\rho = -0.943$.

Table 2: **Estimator ablation.** State matching substantially improves recovery of the continuous oracle-reliance ordering. AUROC and AUPRC evaluate the separate coarse harmful-versus-harmless task.

Estimator	Oracle ρ	AUROC	AUPRC
Unmatched	0.312	0.921	0.796
Unmatched (count-matched)	0.283	0.848	0.633
State-matched	**0.988**	0.840	0.810
State-matched + same-ID subtraction	0.880	0.801	0.667

state matching is necessary for coarse discrimination. Same-identity subtraction provides no demonstrated additional benefit in the evaluated analyses.

5.3 Pair Count Does Not Characterize Support Coverage

Equal pair count can mask substantial differences in both target-space coverage and recovery fidelity (Figure 3). With $N_{\text{cross}} = 300$ fixed, broadly distributed support covers the full target space ($\widehat{\kappa}_Q = 1.0$), whereas concentrated support covers half of it ($\widehat{\kappa}_Q = 0.5$). Concentrated support incurs $12.4\times$ the recovery error of broad support at the median seed, with a seed-level range of 11.7–$14.3\times$. The direction is consistent across all $5/5$ seeds. Thus, N_{cross} measures how many admissible comparisons are available, but not how those comparisons cover the target distribution Q. Pair count therefore cannot replace target-weighted support coverage $\widehat{\kappa}_Q$; the two quantities must be reported separately.

5.4 Stanford EIS: Degradation Without Identified Attribution

Stanford EIS exhibits substantial deployment degradation. Random-holdout MAE ranges from 0.0202 to 0.0331, cell-held-out MAE from 0.0241 to 0.0340, and temporal-holdout MAE from 0.0425 to 0.0576. Relative to the corresponding random-holdout baselines, temporal evaluation increases MAE by 73.4–110.7%.

This degradation does not identify its source. The deployment splits vary multiple aspects of the observed distribution and do not isolate identity at fixed physical state. Stanford therefore shows substantial deployment degradation, but not identified attribution.

Figure 4 illustrates the sensing observations underlying the two real-data evaluations: Stanford EIS spectra and a CWRU vibration window with its corresponding frequency-domain representation.

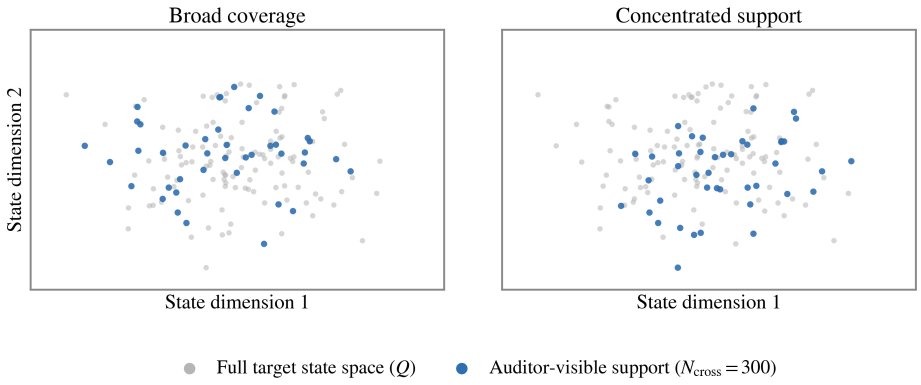
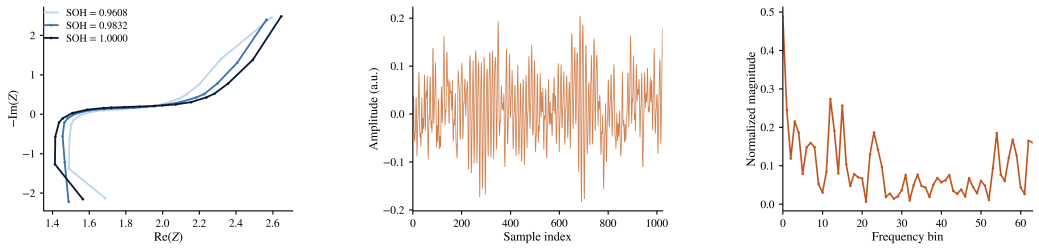


Figure 3: **Equal pair count does not imply equal target-space coverage.** Both audit views contain $N_{\text{cross}} = 300$ admissible cross-identity comparisons but distribute them differently over the target space Q. Broadly distributed support achieves $\widehat{\kappa}_Q = 1.0$, whereas concentrated support achieves $\widehat{\kappa}_Q = 0.5$. Concentrated support incurs $12.4\times$ the recovery error of broad support at the median seed (range 11.7–$14.3\times$; $5/5$ seeds). State-space layouts are schematic.



(a) Stanford battery EIS (b) Raw vibration window (c) 64-bin representation

Figure 4: **Examples of real sensing observations used in the evaluation.** (a) Stanford battery EIS measurements shown in Nyquist coordinates. The displayed spectra are selected deterministically near the 20th, 50th, and 80th SOH percentiles after restricting the reliably labeled population to a common temperature and frequency-grid configuration. (b–c) A CWRU vibration window and its corresponding 64-bin normalized frequency-domain representation produced by the frozen evaluation pipeline. The panels illustrate the sensing observations and do not provide additional attribution evidence.

5.5 CWRU: Support Controls Observational Recovery

Reducing auditor-visible state-region support degrades recovery of the full-support observational reference (Figure 5 and Table 3). As $\widehat{\kappa}_Q$ increases from 0.25 to 1.0, TV error to the full-support reference decreases from 0.227 to 0.127, 0.076, and 0, while rank recovery increases from 0.65 to 0.79, 0.86, and 1.00. Placement instability decreases from 0.299 to 0.164 and 0.100 over the support levels for which it is defined. All three trends are consistent across all five seeds. Because the predictor, predictions, deployment outcome, state and identity definitions, and target distribution Q remain fixed, these differences reflect observational recovery rather than changes in predictor sensitivity.

Under full observational support, predictor sensitivity also retains a deployment-relevant ordering. TV-based sensitivity is positively associated with deployment degradation in all five seeds, with median within-seed Spearman $\rho = 0.785$ and range $[0.685, 0.841]$. Jensen–Shannon divergence preserves the same direction in all five seeds, with median $\rho = 0.832$. The ordering remains positive in all leave-one-condition-out analyses ($60/60$, minimum $\rho = 0.591$) and all leave-one-model-family-out analyses ($20/20$, minimum $\rho = 0.567$). These results establish an observational ordering relationship, not a calibrated causal shortcut effect.

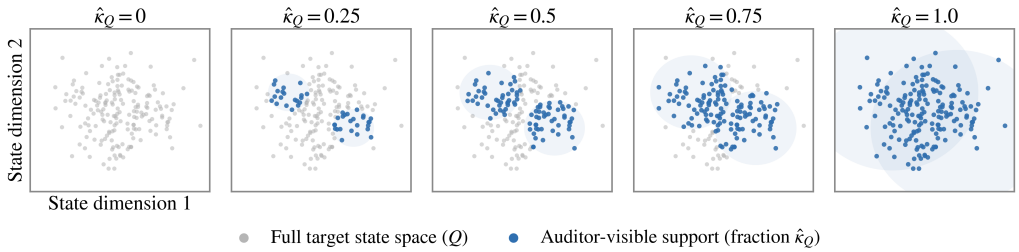


Figure 5: **CWRU support-thinning design.** We vary auditor-visible state-region support while holding the predictor, predictions, deployment outcome, state and identity definitions, and target distribution Q fixed. The state-space layouts are schematic.

Table 3: **CWRU support-thinning recovery.** Increasing target-weighted support improves recovery of the full-support observational reference. All three trends are consistent across all five seeds.

$\widehat{\kappa}_Q$	TV error	Rank recovery	Placement instability
0.25	0.227	0.65	0.299
0.50	0.127	0.79	0.164
0.75	0.076	0.86	0.100
1.00	0	1.00	—

6 Discussion and Limitations

Our results place an identification boundary before shortcut estimation. Predictor-specific reliance compares alternative identities at the same task state, yet observational data may not contain all comparisons required by the target distribution Q. The supported contrast can therefore differ from full-target reliance, and any finite estimate must be interpreted together with its observational support. Partial support reveals predictor behavior only over the comparisons that are observed, while zero support warrants ABSTAIN rather than a zero-reliance conclusion. Pair count is also insufficient because it records how many admissible comparisons exist, not how they cover the target space. Shortcut audits should therefore distinguish full-target reliance, support-restricted reliance, and the finite audit estimate rather than treating them as interchangeable.

The framework remains observational and inherits the limits of its comparison structure. Matching on the measured state description U does not ensure equality of the latent physical state S, so supported cross-identity contrasts should not be interpreted as causal effects. Additional physical, parametric, generative, or invariance assumptions could constrain unsupported comparisons, but identification would then depend on those assumptions rather than on the observational law alone. The controlled study uses a deliberately simple support geometry and a limited set of predictor families, while the real-data evidence covers battery EIS and one mechanically distinct vibration-sensing task. Stanford shows that deployment degradation alone does not identify its source, whereas CWRU provides an observational ordering result rather than a calibrated causal shortcut effect. The framework is therefore intended to quantify predictor sensitivity where comparable evidence exists, characterize the support behind that estimate, and abstain when the required comparison is unavailable.

7 Conclusion

Shortcut auditing from observational sensing data is an identification problem before it is an estimation problem. Our framework distinguishes full-target reliance, support-restricted reliance, and finite audit estimates, and shows how target-weighted observational support governs what can be identified and recovered. Controlled and real-data analyses show that support governs recovery, while deployment degradation alone does not identify attribution. Shortcut audits should therefore report predictor sensitivity together with the support that makes the estimate interpretable, and abstain when the required comparison is unavailable.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing of the manuscript, and to support retrieval and discovery of potentially relevant literature. All AI-assisted writing was reviewed and revised by the authors, and literature suggestions were checked against the original sources before inclusion or citation. The authors take full responsibility for the final manuscript, including its technical claims, analyses, interpretations, and citations.

REPRODUCIBILITY STATEMENT

We provide the assumptions and proof of the identification result in Appendix A. Appendix B specifies the matching procedure, controlled support construction, estimator variants, training seeds, and evaluation pools. Appendix C documents the Stanford and CWRU preprocessing pipelines, predictor reconstruction and validation, support-thinning construction, discrepancy metrics, and statistical conventions. All reported analyses use fixed seeds and predefined evaluation procedures, and undefined zero-support cases are reported as ABSTAIN rather than converted to numerical estimates.

REFERENCES

Eldar D Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems*, 35:17582–17596, 2022.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. Naturalistic causal probing for morpho-syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403, 2023.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, 2(1):31, 2019.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical ai using shortcut testing. *Nature communications*, 14(1):4314, 2023.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Nathan Drenkow, Mitchell Pavlak, Keith Harrigian, Ayah Zirikly, Adarsh Subbaswamy, Mohammad Mehdi Farhangi, Nicholas Petrick, and Mathias Unberath. Detecting dataset bias in medical ai using a generalized and modality agnostic auditing approach. *npj Digital Medicine*, 2026.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, volume 2024, pp. 43832–43868, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3428–3448, 2019.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021.

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. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International conference on machine learning*, pp. 8346–8356. PMLR, 2020.

Gilat Toker, Nitay Calderon, Ohad Amosy, and Roi Reichart. Liberty: A causal framework for benchmarking concept-based explanations of llms with structural counterfactuals. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8149–8186, 2026.

Vibujithan Vigneshwaran, Emma AM Stanley, Raissa Souza, Erik Ohara, Matthias Wilms, and Nils Forkert. Evaluating shortcut utilization in deep learning disease classification through counterfactual analysis. In *medical imaging with deep learning*, 2025.

Ziyao Wang. Martingale doppelgänger-eval: An identification framework for auditing candlestick understanding in vision-language models. *arXiv preprint arXiv:2606.17423*, 2026.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

A Formal Identifiability

A.1 Proof of Proposition 3.1

Part (i): non-identification. We construct two admissible worlds that induce the same observational law but yield different values of full-target reliance. Let $\mathcal{B}$ be a measurable set of requested comparisons with $Q(\mathcal{B}) > 0$ for which the alternative identity branch lies outside observational support. Because the fixed predictor f is nonconstant, there exist $x_0, x_1 \in \mathcal{X}$ such that

$d_Y(f(x_0), f(x_1)) > 0$. Construct two worlds W_0 and W_1 with the same task-state distribution, identity-assignment mechanism, fixed predictor, and observation mechanism on every state–identity pair with positive observational probability. On $\mathcal{B}$, let the observed identity branch produce x_0 in both worlds. Let the unsupported alternative branch produce x_0 in W_0 and x_1 in W_1. The two worlds therefore differ only on an observationally unsupported branch and induce the same observational law,

$$P_{\text{obs}}^{W_0}(X, S, M) = P_{\text{obs}}^{W_1}(X, S, M).$$

The corresponding predictor discrepancies differ on a set with positive Q-mass: the contribution of $\mathcal{B}$ is zero in W_0 and strictly positive in W_1, while all other requested comparisons are held fixed. Hence,

$$R_Q^{W_0}(f) \neq R_Q^{W_1}(f).$$

No functional of the observational law alone can equal $R_Q(f)$ in both worlds. Thus, if Q assigns positive mass to an unsupported requested comparison and no additional structure determines the missing branch, full-target reliance need not be point-identified.

Part (ii): partial identification under bounded discrepancy. Assume $0 \leq D_f \leq D_{\max}$ and $0 < \kappa_Q(\tau) \leq 1$. Conditioning on the support indicator A_τ gives

$$\begin{aligned} R_Q(f) &= \mathbb{E}_Q[D_f(S, M, M')] \\ &= \kappa_Q(\tau)\, \mathbb{E}_Q[D_f \mid A_\tau = 1] + (1 - \kappa_Q(\tau))\, \mathbb{E}_Q[D_f \mid A_\tau = 0] \\ &= \kappa_Q(\tau) R_Q^{\text{sup}}(f; \tau) + (1 - \kappa_Q(\tau))\, \mathbb{E}_Q[D_f \mid A_\tau = 0]. \end{aligned}$$

Because $0 \leq D_f \leq D_{\max}$,

$$0 \leq \mathbb{E}_Q[D_f \mid A_\tau = 0] \leq D_{\max}.$$

Therefore,

$$R_Q(f) \in \left[\kappa_Q(\tau) R_Q^{\text{sup}}(f; \tau),\ \kappa_Q(\tau) R_Q^{\text{sup}}(f; \tau) + (1 - \kappa_Q(\tau)) D_{\max}\right].$$

The interval has width $(1 - \kappa_Q(\tau))D_{\max}$. If $\kappa_Q(\tau) = 1$, the unsupported term vanishes and $R_Q^{\text{sup}}(f; \tau) = R_Q(f)$. If $\kappa_Q(\tau) = 0$, $R_Q^{\text{sup}}(f; \tau)$ is undefined, and boundedness alone gives $R_Q(f) \in [0, D_{\max}]$.

Constructive witness. A finite construction makes the non-identification explicit. Let $S = M = \{0, 1\}$ and $U = S$, and let the target distribution Q place all of its mass on the requested comparison $(S = 1, M = 0, M' = 1)$. Suppose

$$P_{\text{obs}}(M = 1 \mid S = 1) = 0,$$

so the $M = 1$ branch at $S = 1$ is observationally unsupported. Let the observed branch produce a, and consider a fixed predictor satisfying

$$f(a) = 0, \qquad f(b) = 0, \qquad f(c) = 1.$$

Worlds W_0 and W_1 agree on every observationally supported branch. On the unsupported branch $(S = 1, M = 1)$, let W_0 produce b and W_1 produce c. The two worlds therefore share the same observational distribution and the same predictor, but for absolute prediction discrepancy they yield

$$R_Q^{W_0}(f) = 0, \qquad R_Q^{W_1}(f) = 1.$$

The same observational evidence is therefore compatible with different values of full-target reliance.

Scope. The result establishes an observational identification boundary rather than causal attribution. Proposition 3.1 considers the favorable case $U = S$ to isolate the role of observational support. In observational applications, the measured state description U need not fully determine the latent physical state S, so cross-identity matching on U does not by itself isolate a causal identity effect. Physical models, parametric restrictions, validated generative mechanisms, or invariance assumptions may constrain unsupported comparisons and tighten the identified set. Any such identification, however, depends on those additional assumptions rather than on the observational law alone.

B SUPPORT-AWARE AUDIT AND CONTROLLED EVALUATION

B.1 STATE-MATCHED AUDIT

For observations (x_i, u_i, m_i) and state tolerance τ, we define the admissible cross-identity candidate set

$$\mathcal{C}_\tau = \{(i,j) : m_i \neq m_j,\ d_U(u_i, u_j) \leq \tau\}.$$

The implementation uses standardized Euclidean distance over the measured state variables. For a fixed matching seed, the algorithm visits observations according to a fixed seeded permutation. For each unused observation, it selects the nearest unused observation with a different identity and accepts the pair if the state distance does not exceed τ. Ties are resolved deterministically by the lowest remaining eligible index. After a pair is accepted, neither endpoint is reused. We denote the resulting matched set by $\mathcal{M}_\tau^{\text{cross}}$.

The audit reports the number of matched cross-identity comparisons,

$$N_{\text{cross}} = |\mathcal{M}_\tau^{\text{cross}}|,$$

and, when $N_{\text{cross}} > 0$, the state-matched prediction contrast

$$\widehat{R}_\tau^{\text{match}}(f) = \frac{1}{N_{\text{cross}}} \sum_{(i,j)\in\mathcal{M}_\tau^{\text{cross}}} d_Y(f(x_i), f(x_j)).$$

The primary estimator is the unweighted mean over accepted cross-identity pairs. It uses neither same-identity subtraction nor identity-group reweighting.

The audit separately reports empirical target-weighted support coverage $\widehat{\kappa}_Q$. In the controlled design, the target distribution is uniform over K predefined state anchors. If H anchors contain an observable cross-identity comparison, then

$$\widehat{\kappa}_Q = \frac{H}{K}.$$

Under the frozen construction described below, this quantity also equals $2N_{\text{cross}}/N_{\text{audit}}$. The two support statistics therefore measure different properties: N_{cross} records the number of realized comparisons, whereas $\widehat{\kappa}_Q$ records the fraction of the requested target space represented by those comparisons.

If $N_{\text{cross}} = 0$, the numerical reliance estimate is undefined and the audit returns ABSTAIN. The analysis does not encode an undefined comparison as zero reliance. The primary controlled analysis uses $\tau = 0.05$.

B.2 CONTROLLED SUPPORT INTERVENTION

The controlled experiment varies only the cross-identity support exposed to the auditor. For each trained system, we fix the training data, evaluation data, trained predictor, finite full-target oracle reference, deployment degradation, audit-set size, task-state marginal, and identity marginal before constructing any audit view. Exact same-state alternative-identity realizations used to compute the oracle reference are not used for predictor training, observational matching, support-mask construction, or audit-hyperparameter selection.

The audit reservoir contains $K = 20$ state anchors,

$$s_k = \frac{k}{19}, \qquad k = 0, \ldots, 19.$$

Each anchor contains candidate observations from two identities. Every audit view contains 240 observations, with 12 observations per anchor and a balanced identity marginal of $120/120$. We vary the number of anchors with observable cross-identity support as

$$H \in \{0, 2, 4, 6, 10, 16, 20\}, \qquad q = \frac{H}{20}.$$

A supported anchor contributes six observations from each identity. An unsupported anchor contributes twelve observations from one identity. Unsupported-anchor assignments are balanced across

identities, so audit-set size, task-state marginal, and identity marginal remain fixed as support changes.

Adjacent anchors are separated by $1/19 \approx 0.0526$, which exceeds the primary tolerance $\tau = 0.05$. Cross-identity matches can therefore occur only at deliberately supported anchors. Each supported anchor contributes six one-to-one pairs, yielding

$$N_{\text{cross}} = 6H, \qquad \widehat{\kappa}_Q = \frac{H}{20}.$$

At each support level, 20 reproducible masks vary the locations of the supported anchors while preserving H, sample size, per-anchor counts, and both marginals. Mask construction does not use oracle reliance, recovery error, or downstream audit outcomes.

The experiment uses 10 independent training seeds, indexed 42–51, and four predictor families: Ridge, Random Forest, MLP, and a small 1D CNN. Across seven support levels and 20 masks per level, the design contains

$$10 \times 4 \times 20 \times 7 = 5{,}600$$

audit views. Training seed is the independent training repetition. Support masks are repeated observational views of the same frozen system and are not treated as independent replicates.

B.3 Estimator Ablation

We compare four estimators to separate the effects of state comparability, comparison count, and same-identity subtraction.

Auxiliary benchmark worlds. The estimator ablation uses predefined synthetic benchmark worlds that separate identity decodability, predictor reliance, deployment degradation, and observational support. The benchmark includes conditions with decodable identity but no predictor reliance, identity–state association without deployment harm, harmful reliance with finite cross-identity overlap, absent support requiring abstention, task-state extrapolation, and identity shift with the task mechanism retained. These worlds are used only for estimator evaluation and are distinct from the support-only intervention in Appendix B.2.

Unmatched. The unmatched estimator partitions the audit pool by identity, independently permutes the two identity subsets, and pairs observations by position up to the size of the smaller subset. It imposes no state-distance constraint.

Count-matched unmatched. The count-matched unmatched estimator uses the same unmatched construction and then applies a fixed seeded subsample to match the realized pair count of the state-matched estimator. This baseline controls comparison count without enforcing state comparability.

State-matched. The state-matched estimator is $\widehat{R}_\tau^{\text{match}}(f)$ from Appendix B.1. It averages prediction discrepancies over one-to-one cross-identity pairs that satisfy the state-distance tolerance.

State-matched with same-identity subtraction. The subtraction variant removes an analogous state-matched same-identity discrepancy from the state-matched cross-identity discrepancy. It therefore compares between-identity variation against the corresponding within-identity variation.

The primary ablation endpoint is Spearman correlation with the finite oracle-reliance ordering. The oracle-tracking analysis contains 80 defined benchmark-world–model–seed units. AUROC and AUPRC are evaluated separately on a 60-unit harmful-versus-harmless pool. Because the two analyses use different evaluation pools, we interpret them separately. The results support state matching for recovery of the continuous oracle ordering, but they do not establish that state matching is necessary for coarse harmful-versus-harmless discrimination.

C Real-Data Protocols and Reproducibility

C.1 Stanford Battery EIS

The Stanford analysis uses longitudinal battery electrochemical-impedance-spectroscopy measurements with measured state-of-health (SOH) labels. After preprocessing and deduplication, the analysis contains 848 usable spectra from 24 cells, with 87 duplicate spectra removed before evaluation. The prediction target is true SOH. We evaluate six frozen predictors under random, cell-held-out, temporal-held-out, and route-held-out protocols. The random split provides an in-distribution reference. The cell-held-out split evaluates transfer to unseen cells. The temporal split changes longitudinal state coverage together with acquisition-associated structure. The route-held-out protocol is also evaluated, but its behavior is strongly model-dependent and is therefore not summarized by a single aggregate range.

Across the six predictors, random-holdout MAE ranges from 0.0202 to 0.0331, cell-held-out MAE from 0.0241 to 0.0340, and temporal-holdout MAE from 0.0425 to 0.0576. Relative to the corresponding seed-averaged random-holdout baselines, temporal evaluation increases MAE by 73.4–110.7%. These deployment protocols do not independently manipulate the candidate mechanisms and do not provide randomized measurements of the same latent battery state under alternative acquisition identities. The observed degradation therefore does not identify an identity-specific source.

For the Stanford example in Figure 4, we restrict the reliably SOH-labeled population to a common temperature of 25°C and the most frequent measurement-frequency grid. We then select the spectra nearest the 20th, 50th, and 80th SOH percentiles, with deterministic tie-breaking. This selection is used only to illustrate the sensing observations and does not provide additional attribution evidence.

C.2 CWRU Vibration Sensing

The CWRU evaluation changes both the sensing mechanism and the prediction task. The frozen preprocessing pipeline contains 16 vibration measurement files and produces $2{,}400$ processed observations. Each signal is segmented into length-1024 windows with 50% overlap. The pipeline applies a Hanning window, computes the rFFT magnitude, retains the first 64 frequency bins, applies per-window ℓ_2 normalization, and retains at most 150 windows per source file. The final evaluation uses five training seeds, $\{42, 43, 44, 45, 46\}$, and 12 fixed model–target-domain conditions per seed.

Persisted neural-network checkpoints were unavailable, so we reconstruct the neural predictors under the frozen legacy training procedure. Faithful reconstruction requires preserving the original model-outer, seed-inner execution order because model initialization precedes an internal seed reset in the legacy routine. We accept the reconstruction only after it exactly reproduces the independently persisted deployment-degradation values for all 60 final seed–condition units. Predictive-distribution statistics are computed only after this validation succeeds.

The CWRU audit operates on predictive class distributions. Logistic Regression and Random Forest use their native probability outputs, whereas MLP and CNN probabilities are obtained by applying softmax to the reconstructed logits. We apply no post-hoc probability calibration. For a matched pair (i, j), Total Variation is the predefined primary discrepancy,

$$\delta_{ij}^{\mathrm{TV}} = \frac{1}{2} \sum_{k} |p_i(k) - p_j(k)|.$$

Jensen–Shannon divergence with base-2 logarithms is the predefined robustness metric. The full-support matching protocol yields $N_{\mathrm{cross}} = 600$ for every seed–condition unit.

Under full observational support, TV-based sensitivity is positively associated with deployment degradation in all five seeds, with median Spearman $\rho = 0.785$ and range $[0.685, 0.841]$. JS preserves the same direction in all five seeds, with median $\rho = 0.832$. The leave-one-condition-out and leave-one-model-family-out analyses remain positive in $60/60$ and $20/20$ cases, with minimum correlations of 0.591 and 0.567, respectively. These associations characterize observational ordering and are not interpreted as calibrated causal shortcut effects.

For the support-thinning analysis, the target distribution Q is fixed uniformly over four discrete state regions. We expose $r \in \{0, 1, 2, 3, 4\}$ regions to the auditor, giving

$$\widehat{\kappa}_Q = \frac{r}{4} \in \{0, 0.25, 0.5, 0.75, 1\}.$$

The analysis exhaustively enumerates the available region subsets at each nontrivial support level. Predictor parameters, predictions, deployment outcomes, state and identity definitions, Q, and the full-support observational reference remain fixed. Each visible region contributes 150 matched pairs, so $N_{\text{cross}} = 150r$. We evaluate recovery using TV error relative to the full-support reference, recovery of its cross-condition ordering, and variability across supported-region placements. As $\widehat{\kappa}_Q$ increases from 0.25 to 1, TV error decreases from 0.227 to 0.127, 0.076, and 0, while rank recovery increases from 0.65 to 0.79, 0.86, and 1.00. Placement instability decreases from 0.299 to 0.164 and 0.100 over the support levels for which multiple placements exist. All three trends have the same direction across all five seeds.

The fixed-pair-count control separates comparison count from target-space coverage. Both audit views contain $N_{\text{cross}} = 300$ pairs. Concentrated support exposes two of the four state regions and therefore has $\widehat{\kappa}_Q = 0.5$. Broad support subsamples 300 pairs from the full-support pair set while retaining representation from all four regions, giving $\widehat{\kappa}_Q = 1.0$ in the evaluated views. Recovery error is the mean absolute deviation of the thinned-support TV estimate from the full-support observational reference. At the median seed, concentrated support incurs $12.4\times$ the recovery error of broad support, with a seed-level range of 11.7–$14.3\times$ and the same direction in all $5/5$ seeds. Thus, equal N_{cross} does not imply equal target-space coverage or equal recovery fidelity.

C.3 Statistical Conventions and Reproducibility

For the controlled support intervention, training seed is the independent resampling cluster. Model family is a fixed within-seed condition, support level is a within-system intervention, and support mask is a repeated observational view. Endpoint uncertainty is estimated with a cluster bootstrap over the 10 training seeds using $2{,}000$ resamples. Each resampled seed retains all four associated model-family conditions. Support masks are not treated as independent replicates.

For CWRU, training seed is the independent replication unit. The 12 model–target-domain conditions are repeated within each seed. The resulting 60 seed–condition rows are therefore not treated as independent replicates, and the matched pairs within each condition are not treated as independent scientific units.

Whenever no admissible cross-identity pair exists at the specified matching resolution, the reliance estimate is undefined and reported as ABSTAIN with numerical value NA. Undefined estimates are not converted to zero for averaging, correlation analysis, plotting, or harmful-versus-harmless classification. A numerical zero denotes a supported comparison with no observed prediction discrepancy; NA denotes that the requested comparison is unavailable.

All manuscript-final analyses use fixed training seeds and predefined support constructions. The controlled experiment uses no outcome-dependent seed exclusion, support-mask selection, or post-hoc support retuning. The final results use the corrected deterministic one-to-one matcher, and superseded matching outputs are not used for manuscript claims. For CWRU, predictor reconstruction is validated against independently persisted deployment outcomes before any TV or JS analysis is performed. The probability-output path is then checked for hard-label agreement, finite nonnegative probabilities, the expected class dimension, and normalization to one.