THE CONTROL DECIDES THE ANSWER: A MATCHED-OPERATOR NULL FOR EXPLANATION CHANGE UNDER MODEL UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

When a deployed model is updated in response to a distribution shift, its feature attributions move. Existing work measures how far they move and treats the residual, after conditioning on prediction change, as instability. That is not yet a criterion, because it carries no account of how far an explanation *should* have moved. We supply one. Using shift families whose Bayes-optimal predictor is available in closed form, we define a warranted-change reference ω and decompose attribution change across an update into warranted, unwarranted and explainer-noise parts. The design turns on a control prior work does not use: a **matched-operator null**, applying the identical update to fresh *source* data, so treatment and control differ in the distribution alone and not in the amount of training. This changes the conclusion. Across 681 runs the null reproduces a median 82% of the movement that follows an update. On a no-shift placebo, where the answer is fixed by construction, it returns 1.02 on an MLP and 0.98 on gradient-boosted trees with exact TreeSHAP, while the seed-retraining floor prior work uses implicitly returns 0.32 and 1.90 on the same data, disagreeing with itself sixfold. We state the consequence for our own results instead of leaving it to be found: against that seed floor our headline effect is inverted (median 0.31), so the case for the matched null rests on the placebo. Applying the missing control to a published audit, 64% of its 45 settings produce movement smaller than refitting on a bootstrap resample of the same data. Unwarranted change is real but bounded: present across MLPs over a 609$\times$ parameter range, trees and a small ResNet, weaker on real census covariates, absent on DistilBERT. Magnitude indicates legitimacy only once an update completes the adaptation it implies, so it fails in the light-update regime auditors work in.

1 INTRODUCTION

Post-hoc feature attributions are consumed as evidence: a practitioner reads which features a model relied on and acts on what they read. Models do not stand still. They are fine-tuned when data drifts, retrained on refreshed samples and redeployed. Whether the evidence survives such an update is an open question.

The question has been asked in pieces. Mougan et al. (2023) compare explanation *distributions* across domains for a frozen model. Subramaniakuppusamy & Gajjar (2026) filter to prediction-preserved pairs before measuring attribution stability under input perturbation. Hemmat & Fatemi (2025) difference attributions across a model update. Dhayalkar (2026) tracks attribution change across fine-tuning epochs. Each measures a magnitude. None supplies the prior quantity, which is how far the explanation should have moved. Absent that quantity, every method treats change as damage and names the residual by fiat, calling it instability, semantic drift or risky reliance redistribution.

Treating all change as damage misclassifies a case that occurs in practice. A model fine-tuned on data from which a spurious shortcut has been removed should stop attributing to that shortcut. Its predictions on most inputs are unchanged, its explanation changes a great deal, and the change is the

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.

desired behaviour. A metric that scores this as instability is measuring something other than what it claims.

Two checkpoints that agree on the shared support but differ in their gradients are different functions that the data did not distinguish, which is the underspecification phenomenon (D'Amour et al., 2022) read through attributions. There is a second gap, and it proves to be the larger one. Measuring how far attributions moved after an update requires knowing how far they would have moved anyway. Prior work's implicit control is an independent retrain with a different seed. That control differs from the treatment in operator, data volume and step count at once, so a comparison against it does not isolate the distribution. We supply a control that differs in the distribution alone, and find that the choice reverses the ordering of update regimes.

Contributions.

1. **A warranted-change reference.** We define ω as the change in the attributions of the *Bayes-optimal predictor* of each environment, not of the causal generating mechanism. On generators where that predictor is available in closed form, ω is exact: identically zero under covariate shift, non-zero and computable under concept shift and shortcut removal. The definitional choice is what makes the shortcut case come out right (§3).
2. **A matched-operator null.** The control for a fine-tune-on-target treatment is the identical fine-tune on fresh *source* data, matched in learning rate, epochs, step count and sample size (§3, §6.2).
3. **A decomposition, not a metric.** Every distance we use already exists (Alvarez-Melis & Jaakkola, 2018; Yeh et al., 2019; Bhatt et al., 2020; Agarwal et al., 2022a;b; Hedström et al., 2023); the contribution is the reference and the controls.
4. **Three propositions** on which explainer classes inherit stability from prediction stability, each with a sharpness construction and a per-point numerical check (§4).
5. **An empirical finding that survives its own controls**, together with an explicit account of where it stops holding (§6).

What we do not claim. We do not propose a new stability metric. We do not detect distribution shift; that is the contribution of Mougan et al. (2023), whose frozen-model setting is not ours. We were not first to condition on prediction preservation; Subramaniakuppusamy & Gajjar (2026) were, and we adopt their conditioning. Nor do we use the phrase "explanation drift", which denotes at least three distinct objects in the recent literature (Appendix A).

2 Setup and related work

A category error worth stating once. For a frozen model f and a deterministic explainer E whose baseline and background sample are held fixed, $E_f(x)$ is a fixed function of x: changing $P(X)$ changes only which x are observed. Instance-level change under pure data shift is then identically zero, and the only measurable object is the shift of the explanation *distribution*, which is what Mougan et al. (2023) measure. The proviso matters: expected gradients, KernelSHAP and LIME average over a background set, which we fix once and reuse at every checkpoint, since refreshing it to the shifted $P(X)$ moves $E_f(x)$ with f frozen. Genuine instance-level change otherwise requires that the model changed, this paper's case; that the input changed, the concern of Subramaniakuppusamy & Gajjar (2026) and the adversarial-fragility line (Ghorbani et al., 2019; Dombrowski et al., 2019; Slack et al., 2020); or that the explainer is stochastic, a nuisance we measure and subtract. We verify all three empirically instead of asserting them (§6.1).

Objects. $f : \mathcal{X} \to \mathbb{R}$ is the explained scalar output, always the logit, so that theory, explainers and reference share a space; $p = \sigma(f)$ is the induced probability. An update operator produces $f_{t+1} = \mathcal{U}(f_t, \mathcal{D})$. A probe set $\mathcal{P}$ is drawn from the *shared support* of source and target, so neither checkpoint extrapolates on it, and the prediction-preserved subset is $\mathcal{P}_\epsilon = \{x \in \mathcal{P} : |p_t(x) - p_{t+1}(x)| \leq \epsilon\}$, $\epsilon = 0.05$. Preservation is thus constrained in probability space, whereas Proposition 1 bounds attribution change in logit space, at x and at the baseline. Since σ' vanishes at saturation the one does not bound the other, and on $\mathcal{P}_\epsilon$ the logit gap runs to a median of 0.17 and a maximum of 2.36 on

Table 1: Related work organised by what changes. The last row is closest to this paper.

What changes	Work	What is missing
data, model frozen	Mougan et al. (2023)	instance-level change is vacuous here
the input	Subramaniakuppusamy & Gajjar (2026); Ghorbani et al. (2019)	single model; residual named by fiat
the seed	Bensmail (2025); Laberge et al. (2023)	no shift; this is our *floor*
the model class	Thackshanaramana B. (2026)	architecture varies, distribution does not
the checkpoint	Hemmat & Fatemi (2025); Dhayalkar (2026); Elangovan & Ting (2026)	no reference, no matched control

the generator, pooled over the preserved points of 12 runs (`measure_logit_gap.py`; the per-run median of medians is 0.20). Conditioning on preservation is not optional: Shapley efficiency implies that if two inputs' predictions differ their attribution vectors must differ in at least one coordinate, so measuring change without conditioning partly measures an algebraic necessity. Shifts are *generated*, never assumed. Covariate shift moves $P(X)$ with $P(Y \mid X)$ fixed, concept shift moves $P(Y \mid X)$, and shortcut removal severs a feature that was predictive but never causal.

A separate line explains the shift itself instead of the model's response to it (Kulinski & Inouye, 2023; Hinder et al., 2023), and a further line documents that explainers disagree with one another on a fixed model (Krishna et al., 2022). Neither supplies a reference for how far an explanation should move. We exclude attention weights, whose status as explanation is contested (Jain & Wallace, 2019). Table 1 organises the field by the object that changes; the two nearest deserve specifics. Hemmat & Fatemi (2025) difference occlusion attributions across model pairs and label movement uncorrelated with behaviour change "spurious", which is a heuristic and not a reference, and which misfires on the shortcut case. Subramaniakuppusamy & Gajjar (2026) establish that without conditioning on prediction preservation up to 99% of compared pairs have changed predictions, then measure raw distance on what remains, with no floor for comparison.

3 Warranted change and the two controls

The reference. We define ω against the Bayes-optimal predictor $h_e(x) = \operatorname{logit} P_e(Y{=}1 \mid X{=}x)$ of each environment, $e \in \{S, T\}$ indexing source and target:

$$\omega(x) = d\big(\phi(\mathrm{IG}[h_S](x, b)),\ \phi(\mathrm{IG}[h_T](x, b))\big), \tag{1}$$

where IG is integrated gradients (Sundararajan et al., 2017), $\phi(a) = |a|/\|a\|_1$ maps attributions to the simplex and returns zeros on an all-zero row, $d(u, v) = \frac{1}{2}\|u - v\|_1$, and the baseline b is fixed across checkpoints. The measured change is the same functional at the two checkpoints, $\Delta(x) = d(\phi(E_{f_t}(x)), \phi(E_{f_{t+1}}(x)))$. A causal reference would invert the shortcut verdict: under shortcut removal the causal function is unchanged, yet a correct model must stop using the shortcut. The Bayes reference is exactly zero under covariate shift and available in closed form under concept shift and shortcut removal.

The floors. ν is the explainer's own noise, measured between two runs on one checkpoint. ρ_{null} is the matched-operator null: the identical update applied to fresh source data at matched learning rate, epochs, steps and sample size, so only the sampling distribution differs. ρ_{seed} is the independent-retrain floor that prior work uses. Every term below is a mean over $\mathcal{P}_\epsilon$:

$$\mathrm{UEC} = \mathbb{E}_{\mathcal{P}_\epsilon}[\Delta] - \mathbb{E}_{\mathcal{P}_\epsilon}[\omega] - \max\big(\mathbb{E}_{\mathcal{P}_\epsilon}[\nu], \mathbb{E}_{\mathcal{P}_\epsilon}[\rho_{\mathrm{null}}]\big), \quad \mathrm{ratio} = \mathbb{E}_{\mathcal{P}_\epsilon}[\Delta]/\mathbb{E}_{\mathcal{P}_\epsilon}[\rho_{\mathrm{null}}]. \tag{2}$$

UEC is a difference, not a ratio to output change (Agarwal et al., 2022a), which is undefined and explosive in the prediction-preserving regime this paper concerns. The null and the treatment contain the same optimisation transient, and only the treatment additionally contains the distribution signal.

4 Theory

Let $\delta = f_t - f_{t+1}$ and $\gamma(\alpha) = b + \alpha(x - b)$. All explainers below are linear in the explained function. Completeness is stated against a reference: a fixed baseline b for IG, an expectation over

a background sample $\mathcal{B}$ for KernelSHAP and expected gradients. Throughout this section ϵ is a tolerance on δ and so lives in *logit* space; it is a free parameter of each statement, not the probability threshold $\epsilon = 0.05$ that defines $\mathcal{P}_\epsilon$ in §2. The two are not interchangeable, which is the point of the preceding paragraph and the reason §6.6 measures the logit gap rather than inferring it.

Proposition 1 (Path-integrated: the aggregate is inherited, the allocation is not)**.** *(i)* $\sum_j \Delta\mathrm{IG}_j(x) = \delta(x) - \delta(b)$ *exactly, hence* $|\sum_j \Delta\mathrm{IG}_j(x)| \leq 2\epsilon$ *when the checkpoints agree to* ϵ *at* x *and at* b*, requiring agreement at the two endpoints only and not along the path. (ii)* $\|\Delta\mathrm{IG}(x)\|_1 \leq \|x - b\|_1 \sup_\gamma \|\nabla\delta\|_\infty$*. (iii)* Sharpness: *output preservation alone does not bound individual components. With* $\delta(z) = \epsilon \sin(k(z_1 - z_2))$*,* $b = \mathbf{0}$*,* $x = \mathbf{1}$*, we have* $|\delta| \leq \epsilon$ *everywhere and the aggregate is zero, yet* $\Delta\mathrm{IG}_1 = \epsilon k$ *is unbounded.*

Remark. Part (i) needs agreement at b as much as at x, which $\mathcal{P}_\epsilon$ does not supply: $\mathbf{0}$ is the mode of the data distribution but lies outside its typical set, where no draw falls. We measure $\delta(b)$ rather than assume it.

Proposition 2 (Local gradients inherit nothing, not even the aggregate)**.** *Take* $\delta(z) = \epsilon \sin(kz_1)$ *with* $k = 2\pi n / x_1$ *for integer* n*. Then* $|\delta| \leq \epsilon$ *everywhere and* $\sum_j \Delta\mathrm{IG}_j(x) = 0$*, yet* $\sum_j \Delta(\mathrm{G}\times\mathrm{I})_j(x) = x_1 \epsilon k \cos(kx_1) = 2\pi n\epsilon$ *is unbounded in* n*.*

Proposition 3 (Shapley, under a coalition-level premise)**.** *If* $|v_f(S) - v_{f'}(S)| \leq \epsilon_{\mathrm{coal}}$ *for every coalition* S*, then* $\|\varphi^f - \varphi^{f'}\|_\infty \leq 2\epsilon_{\mathrm{coal}}$*.*

Remark. ϵ_{coal} is a supremum over masked, off-manifold inputs, on which fine-tuning places no constraint, so preservation on the data implies nothing about it. We measure the gap (§6.6).

Corollary. Efficiency gives Shapley the aggregate identity completeness gives IG: $\sum_j \Delta\varphi_j(x) = \delta(x) - \delta(b)$ against a fixed baseline, $\delta(x) - \mathbb{E}_{\mathcal{B}}[\delta]$ against a background distribution. Neither needs the coalition premise, which Proposition 3 requires only for the allocation; and of the two, $\mathbb{E}_{\mathcal{B}}[\delta]$ is the better controlled, averaging over draws from $P(X)$ as $\delta(b)$ does not. The split that survives is completeness, not class: saliency, gradient×input, SmoothGrad and LIME state no aggregate identity and inherit nothing, whereas IG, expected gradients and KernelSHAP inherit the aggregate and nothing beyond. None is protected in its *allocation* across features, the part a practitioner reads: Proposition 1(iii) breaks IG's, and Shapley's rests on a premise we measure to fail. Collinearity constrains it further, no ranking being simultaneously faithful, stable and complete when features are collinear (Caraker et al., 2026); our redundant-feature block is built to exercise that case. Nor is the construction exotic: a difference that oscillates transverse to the integration path while vanishing on it is two checkpoints that agree on the data manifold and curve differently off it, the mechanism Dombrowski et al. (2019) identified in input space, transposed to model space.

5 Experimental protocol

Shifts are generated from a structural equation model with d=20 features in causal, shortcut, redundant and noise blocks, for which the Bayes-optimal logit and its integrated gradients have closed forms. Attributions target the logit; a per-checkpoint attribution baseline would make ΔIG uninterpretable. Activations are smooth by default, because on a ReLU network IG completeness holds only to $O(1/n_{\mathrm{steps}})$ and would contaminate the Proposition 1 check. Gradient-boosted trees use XGBoost with exact TreeSHAP (Chen & Guestrin, 2016). We use seven explainers spanning path-integrated, local-gradient and Shapley-type classes (Sundararajan et al., 2017; Erion et al., 2021; Simonyan et al., 2013; Smilkov et al., 2017; Lundberg & Lee, 2017; Ribeiro et al., 2016; Lundberg et al., 2020), implemented through Captum (Kokhlikyan et al., 2020). Beyond the generator we use ACS Income (Ding et al., 2021), CIFAR-10 (Krizhevsky, 2009) under corruptions we generate locally at mild severity from the families of Hendrycks & Dietterich (2019), and DistilBERT on IMDB (Sanh et al., 2019; Maas et al., 2011). Tests are paired across seeds with Holm correction (Holm, 1979) and Cliff's δ (Cliff, 1993). Our reading of what an attribution evaluation must control for follows recent critiques of the evaluation protocols themselves (Kiourti et al., 2025; Sridhar et al., 2026). Figures are drawn to be legible under the common colour-vision deficiencies and in monochrome (Appendix J).

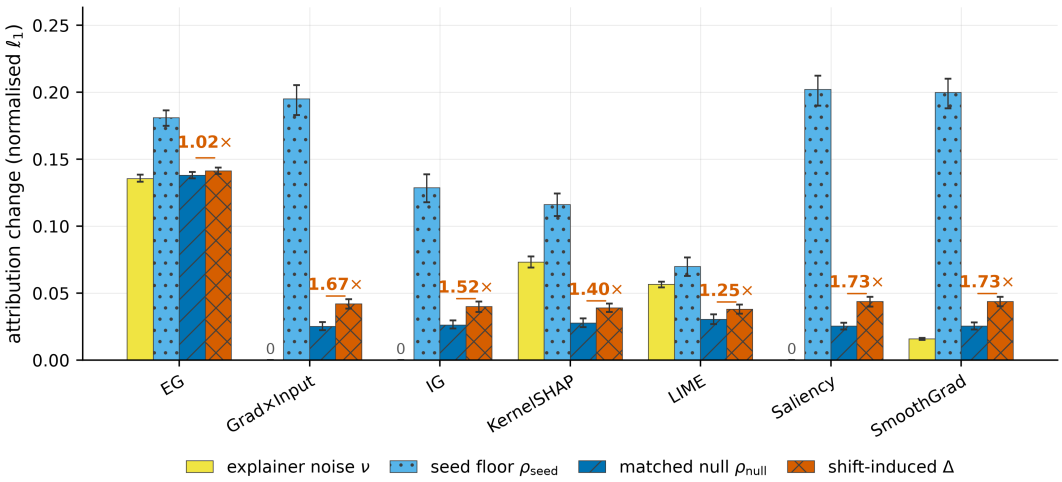


Figure 1: Attribution change against its floors under covariate shift, where ω is exactly zero, so any movement above the matched null is unwarranted. Bars are means over seeds with bootstrap intervals; brackets give $\Delta/\rho_{\text{null}}$. Deterministic explainers have $\nu = 0$ exactly, marked "0". The seed floor ρ_{seed} exceeds the measured change for every explainer, which is the observation §6.2 turns on.

6 Results

6.1 The controls behave, and the placebo is decisive

Deterministic explainers reproduce bitwise on a fixed checkpoint, so $\nu = 0$ exactly; a frozen model under pure data shift yields identically zero instance-level change, confirming §2 numerically and not by assertion. On a no-shift placebo, where the correct answer is fixed in advance, the matched null returns 1.02 on an MLP and 0.98 on gradient-boosted trees with exact TreeSHAP. The seed floor returns 0.32 and 1.90 on the same data. It disagrees with itself sixfold across two model classes and errs in opposite directions, so no fixed correction recovers it. This test does not depend on any shift result and was specified before them.

6.2 Which control you choose reverses the conclusion

Over a grid of shift magnitude $\times$ update strength (600 runs, 5 seeds), change relative to the matched null *falls* as the update gets heavier, and change relative to the seed floor *rises* (Figure 2). For IG at shift magnitude 1.5 and update learning rate 2×10^{-4}, $\Delta/\rho_{\text{null}}$ runs 1.81, 1.58, 1.39, 1.28, 1.09 across 1, 2, 5, 20 and 60 epochs; $\Delta/\rho_{\text{seed}}$ runs 0.11, 0.19, 0.41, 0.82, 1.07. That is the most pronounced of the twelve cells, so we report what happens in the rest: the opposite ordering holds in 24 of 24 explainer $\times$ magnitude $\times$ learning-rate cells, and the curves cross in 10 of 12. The direction, which is what the argument needs, is unanimous. Read against the seed floor a heavy update looks maximally shift-driven; read against the matched null it is the least shift-driven condition. The reason is mechanical: the seed floor is a fixed quantity and Δ grows with training, so the ratio to it measures how hard one trained, not how far the data moved. Across 681 runs the matched null reproduces a median 82% of the movement that follows an update.

Stated plainly: against the seed floor our headline effect is inverted. In the covariate-shift condition of §6.3, $\Delta/\rho_{\text{seed}}$ runs 0.21 to 0.78 with a median of 0.31. A reader who takes seed retraining as the reference will conclude there is no phenomenon here. We put the number in the main text because it is the first thing a sceptical reader computes, and because our answer to it is the paper's argument and not a qualification of it. The obvious suspicion is that we adopted whichever control produced the answer we wanted. The placebo of §6.1 speaks against that, and it is decisive because its ground truth is fixed in advance: a control that manufactures a two-fold effect from an unshifted placebo cannot certify effects under shift. Two further points support it. The comparison is not like-for-like, since ρ_{seed} retrains from scratch on 8,000 points for 60 epochs, 1,920 steps,

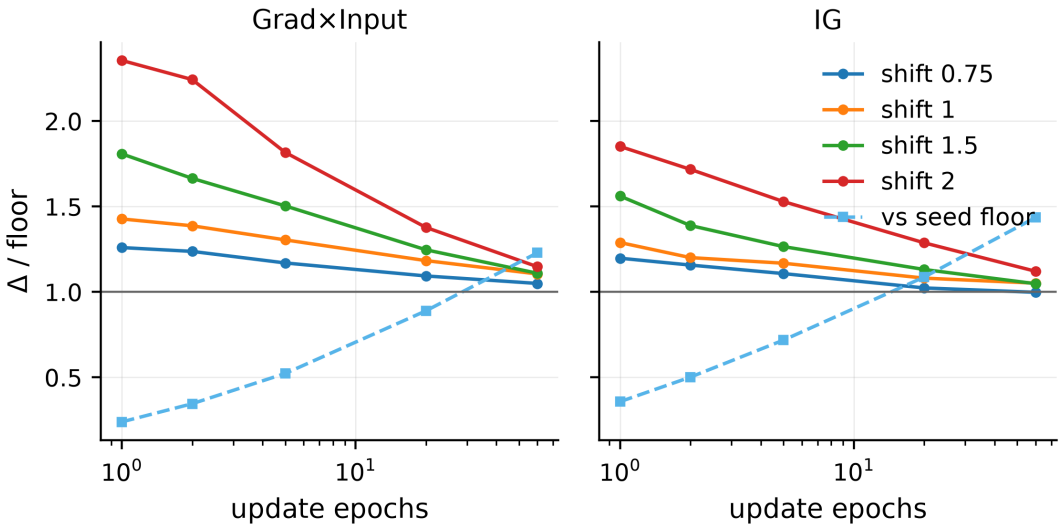


Figure 2: Change relative to the matched null (solid) and to the seed floor (dashed) as the update grows heavier. The two controls order the update regimes in opposite directions and the curves cross. Learning rates are pooled here; the single cell quoted in the text is more pronounced.

whereas the update is 2 epochs on 4,000, 32 steps, so $\Delta/\rho_{\text{seed}}$ varies operator, data volume, learning rate and step count at once. And the two controls disagree about direction, not only magnitude; a valid-but-conservative reference would preserve the ordering and change only the scale.

6.3 Unwarranted change exists

Under covariate shift, where ω is exactly zero, all seven explainers move more than their matched null (Figure 1). The ratios are 1.73 (saliency and SmoothGrad), 1.67 (gradient×input), 1.52 (IG), 1.40 (KernelSHAP), 1.25 (LIME) and 1.02 (expected gradients), with Holm-adjusted $p = 0.014$ and Cliff's δ up to $+0.98$. The effect is largest for gradient-boosted trees with exact TreeSHAP (2.08), so it is not a gradient artefact, and it persists on a small CIFAR ResNet at 1.37. That arm also answers whether the shortcut result is an artefact of a token feature: replacing the six-pixel corner patch with a class-correlated *background*, the geometry a natural spurious correlation has, returns 1.33 against the patch's 1.34 and with less seed variance (Appendix H). The same arm shows the measure discriminating rather than condemning all movement, since a pure brightness shift sits at its own null (1.03, with 6.4% of probe points above the null's 95th percentile against the 5% the floor defines). It also replicates on ACS Income, where 27 of 28 explainer × state cells remain significant after Holm correction. Ablating the distance, normaliser, ϵ threshold, IG baseline, SHAP background and probe sampling leaves the ordering of conclusions unchanged (Kendall $\tau = 1.00$ across the core axes).

6.4 Magnitude and legitimacy: a conditional result

We now test whether the amount an explanation moves indicates whether it should have moved. Within a shift family each probe point carries its own ω, so measured change can be regressed on it directly. Correlations are computed within seed and aggregated across seeds. Under light updates the association is small: $r = +0.121$ $[+0.084, +0.157]$ under concept shift over 3,237 points and $r = -0.104$ $[-0.192, -0.015]$ under shortcut removal over 4,329 points, so the warranted change accounts for roughly 1% of the variance in the measured change. Both intervals exclude zero, so we do not claim the association is absent, only that it is small. We report this as a magnitude with an interval instead of testing against an equivalence margin, since any margin chosen after seeing these numbers would be the post-hoc move this paper criticises elsewhere.

The per-explainer version of the same comparison is underpowered at ten seeds. Two one-sided tests (TOST) (Schuirmann, 1987) against a ± 0.10 margin, fixed from the placebo's own spread, declare

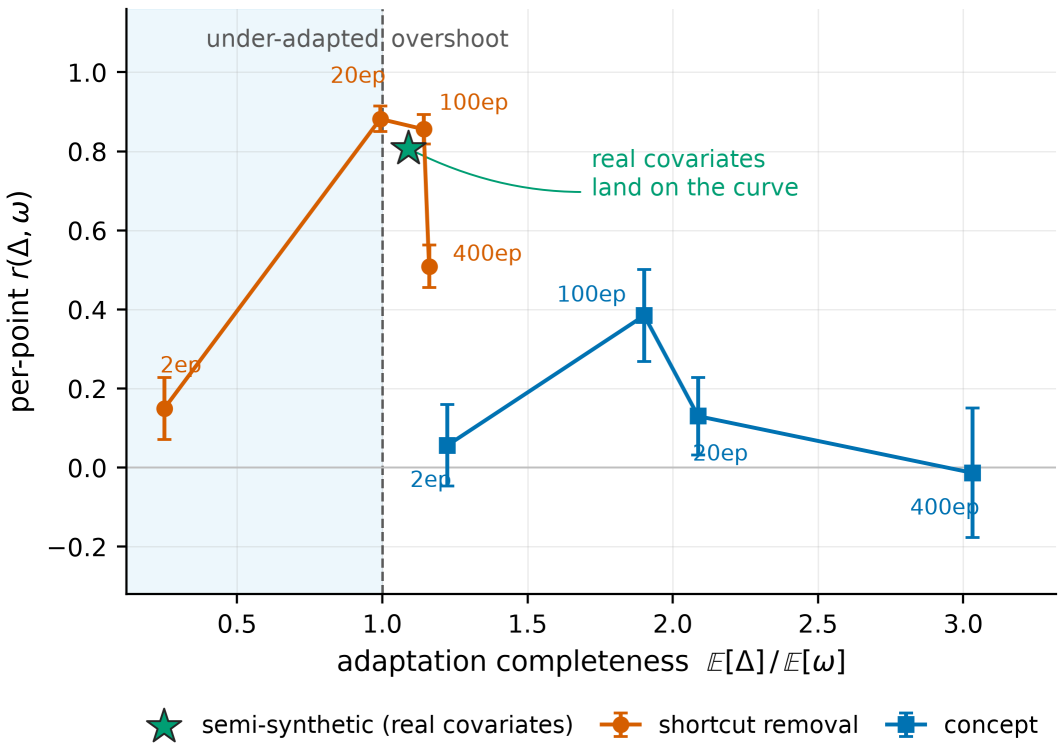


Figure 3: Per-point tracking of the warranted reference against how far the update actually travelled. Sweeping the update budget carries r from $+0.15$ to $+0.88$ as completeness passes through 1, then down again on overshoot. The semi-synthetic run on real covariates lands on the same curve, so the disagreement in §6.4 is not a property of the data.

only one of seven explainers equivalent, and that one is expected gradients, whose two ratios are both near 1.02 and whose pass therefore reports its insensitivity. LIME is a genuine exception: its ratio is 0.64 higher under shortcut removal than under covariate shift ($p = 0.002$). We name it instead of averaging it away, while noting that its own noise floor exceeds its matched null, so we cannot separate sensitivity from sampling variance.

The condition. Repeating the measurement on real ACS covariates with a mechanism known by construction gives $r = +0.81$, the opposite conclusion. The cause is not the covariates. Define completeness as $\mathbb{E}[\Delta]/\mathbb{E}[\omega]$, every expectation here being over $\mathcal{P}_\epsilon$ as in Equation 2: the fraction of the warranted distance the update actually covered; it is 0.12 and 0.54 in the two synthetic settings and 1.09 in the semi-synthetic one. Sweeping the update budget on the generator confirms the reading (Figure 3): on shortcut removal, r rises from $+0.149$ at completeness 0.249 to $+0.882$ at completeness 0.993, then falls to $+0.509$ once the update overshoots. Distance from perfect adaptation predicts tracking at $\mathrm{corr}(|\log \text{completeness}|, r) = -0.870$, and on the rising limb completeness and r correlate at $+0.756$. Because $\mathbb{E}[\Delta]$ contains the optimiser transient, we also report the null-adjusted numerator: on the same rows $(\mathbb{E}[\Delta] - \mathbb{E}[\rho_{\text{null}}])/\mathbb{E}[\omega]$ moves the rising-limb correlation only to $+0.714$ and strengthens the distance-from-perfect one to -0.895, so the reading does not rest on leaving the transient in. The relationship is peaked and not monotone, which is why a naive pooled correlation of -0.356 would mislead. The claim is therefore conditional: attribution magnitude tracks warranted change when the update completes the adaptation, and under the light, prediction-preserving updates that deployment uses it carries little such information. That is the regime auditors work in, because heavy updates which complete the adaptation also move predictions visibly.

Two published metrics inherit the consequence. Reimplemented on the same checkpoints, a FASS-style filtered distance and Delta-Audit's spurious residual score correct adaptation as 45% and 89% *worse* than unwarranted drift, and UEC assigns them opposite signs.

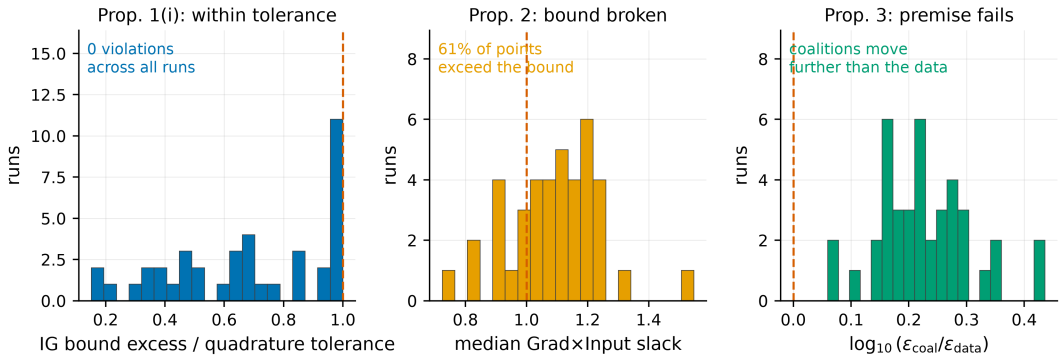


Figure 4: Per-run checks of the three propositions. Left: IG's aggregate excess over the bound, divided by each run's own quadrature tolerance, so values below 1 are numerical. Middle: gradient×input breaks the same bound. Right: the coalition premise of Proposition 3 fails off-manifold.

6.5 Re-auditing a published result

Applying the missing control to Hemmat & Fatemi (2025), we re-run its 45 reported settings with a matched null. 64% produce attribution movement smaller than refitting the same model on a bootstrap resample of the same data, drawn with replacement at the original sample size, and only 33% clear that floor, with a median ratio of 0.861. That floor is the lower of the two on offer: at matched sample size a bootstrap refit moves a median 0.75 of the distance a refit on a disjoint fresh sample moves, in 17 of 18 learner × dataset cells, so the 64% understates what a fresh-draw floor would give (Appendix F). Their own flagship examples survive comfortably, which is worth stating: the conclusion is about the reported settings in aggregate and not about the paper's headline claims.

6.6 The theory holds, and the change is not faithfulness loss

Proposition 1(i) records 0 violations in 20,000 probe points, once each run is compared against its own quadrature tolerance (Figure 4); gradient×input breaks the same bound on 60.8% of points. The coalition premise of Proposition 3 fails off-manifold, with masked inputs moving 1.49 to 2.02 times further than data inputs across shift families. Separately, the effect is not explainers becoming wrong instead of merely different: restricting to points where the explainer is faithful to both checkpoints leaves the ratio unchanged (1.567 against 1.563), and the correlation between Δ and the faithfulness drop lies in $[-0.084, +0.032]$ across seed-averaged cells.

6.7 Where the effect stops

The effect is bounded, and we map the boundary and not the interior. It is flat at roughly 1.4 across a 609× parameter range within models trained from scratch, moving from 1.48 at 1.8k parameters to 1.36 at 1.07M. On real census covariates with a mechanism known by construction it weakens to 1.20 $[1.05, 1.34]$ for IG, and the machinery of ω transfers exactly there, matching quadrature to 1.1×10^{-14} on real rows under a tilt whose domain-classifier AUC, 0.901, matches the generator's at magnitude 0.75, milder than the 1.5 used above (Appendix E). On DistilBERT at 66.9M parameters it is absent: 1.02 $[0.97, 1.09]$, unchanged across a tenfold range of update strength.

Modelling the optimiser share directly shows that the last result is not simple extrapolation (Figure 5, Appendix B). Capacity does not predict the share at all, with a logit-scale slope of -0.007 $[-0.197, +0.184]$, $p = 0.94$, across 609× of parameters. Update strength does predict it, monotonically. At matched update size DistilBERT sits $+0.173$ $[+0.104, +0.243]$ *above* the from-scratch curve (Welch $p = 3.3 \times 10^{-4}$). Whether that excess comes from pretraining or from architecture, this design cannot separate.

Finally, the monitored quantity that does correlate with unwarranted change points the wrong way: higher prediction agreement between checkpoints goes with *more* unwarranted change, not less, and target accuracy's association reverses sign across shift magnitudes.

7 LIMITATIONS

The reference needs a known Bayes predictor. ω is exact only where the generative process is known. On real data we claim the covariate-shift null alone, and even there it is checked through a calibration-transfer test and is not proved. The semi-synthetic construction removes the covariate half of that assumption but not the label half, which no method can remove.

Magnitude tracks legitimacy only when the update completes the adaptation. We identify the governing quantity (§6.4) but characterise it on two shift families in one generator plus one semi-synthetic setting. Whether completeness governs tracking elsewhere is untested.

Scale. DistilBERT is covered, but nothing here speaks to vision transformers, and the $+0.173$ excess cannot be attributed to pretraining over architecture without a same-size transformer trained from scratch.

Naturally-spurious benchmarks. ω needs a Bayes-optimal predictor in closed form, which Waterbirds, Camelyon17 and ImageNet-C do not supply. The matched null transfers to them; the warranted reference does not. Half the contribution is therefore untested at high dimension. What we can do is give the null a confound shaped like theirs, and Appendix H does: a class-correlated background at 3072 dimensions reproduces the corner-patch result. That is evidence the geometry does not matter, not evidence that ω scales.

The two measurements are orthogonal, and we measured that rather than asserting it. Reimplementing the explanation-distribution detector of Mougan et al. (2023) and running it on our checkpoints (Appendix G) gives 0.956 on its own object, covariate shift under a frozen model, and chance on concept shift and shortcut removal there, because a frozen model's attributions cannot respond to a change in $P(Y \mid X)$. Turned on a model update it stays within 0.06 of its own floor. Ours is identically zero in their setting and 1.07 to 2.24 in ours. Neither instrument subsumes the other, which is what §2 claimed on structural grounds.

An exponential moving average smooths, it does not reference. We ran the EMA arm rather than argue for it, in the form PyTorch implements and in a bias-corrected one (Appendix I). Judged at its best, corrected, EMA read *as* the reference still fails the placebo: it returns 1.25 to 1.72 across decays where the answer is 1 by construction and the matched null returns 1.08. Uncorrected it fails far worse and without bound, reaching 54.0 at decay 0.999, because an average seeded at the starting weights still carries 96.8% of them after the 32 steps of this update and so largely reports the source model. Applied to *both* arms with the matched null still in place it returns 1.10 and leaves the ordering intact. The two are complementary, and only one of them is a control.

Our effect is smaller than the literature we critique. Against the seed floor it is inverted, and the case for the matched null rests on a placebo rather than on the shift results the control is used to interpret. A reader who rejects the placebo argument should reject the paper's empirical claims with it.

8 CONCLUSION

The measurement of explanation change after a model update turns on a control that has not been used. Applying the identical update to fresh source data isolates the distribution from the optimisation, and it reverses the ordering that a seed-retraining floor produces. Most of the attribution movement reported after fine-tuning is reproduced by the optimiser alone. A reference for how far explanations should have moved makes the remaining question askable, and answers it conditionally: magnitude indicates legitimacy when an update completes the adaptation it implies, and carries little information under the light updates that deployment uses.

ETHICS STATEMENT

This work uses only public datasets and involves no human subjects, no personally identifying information and no new data collection. ACS Income (Ding et al., 2021) is derived from public census microdata and carries demographic attributes including sex and race, which our models consume as features because the published prediction task does. We make no fairness claim and use the data solely to test whether a measurement design transfers from a generator to real covariates; the attri-

bution changes we report should not be read as evidence about any group. The paper is diagnostic in aim, and the plausible misuse we can foresee is that a practitioner reads unwarranted attribution change as a fairness finding, which §6.4 argues directly against. The authors have read the ICLR Code of Ethics and adhere to it.

Reproducibility statement

Every quantitative claim is regenerated from the released code. A single entry point runs all fourteen stages in dependency order and rebuilds every table and figure. A verification script re-derives 81 headline numbers from the produced artefacts and fails if any drifts, and 101 tests cover the generator, the operators and a numerical check of each proposition. Figures write their source data beside the image. The synthetic and semi-synthetic pipelines run on CPU; the transformer arm requires one GPU and its measured output is committed so the rest reproduces without it. Code is supplied as anonymised supplementary material.

Use of AI statement

Large language models were used as a coding and editing assistant: drafting and refactoring experiment code, proposing figure layouts, and copy-editing prose. All experimental design decisions, the choice of reference and controls, the propositions and their proofs, and every claim in this paper were made and checked by the authors. Every number reported here is produced by the released code and re-derived by the verification script described above.

References

Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations. *arXiv preprint arXiv:2203.06877*, 2022a. ICLR 2022 PAIR^{2}Struct workshop.

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2022b.

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Chama Bensmail. EvoXplain: When ML models agree on predictions but disagree on why. *arXiv preprint arXiv:2512.22240*, 2025.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Drake Caraker, Bryan Arnold, and David Rhoads. The attribution impossibility: No feature ranking is faithful, stable, and complete under collinearity. *arXiv preprint arXiv:2605.21492*, 2026.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Sahil Rajesh Dhayalkar. Reasoning stabilization point: A training-time signal for stable evidence and shortcut reliance. *arXiv preprint arXiv:2601.11625*, 2026. Submitted to ACL Rolling Review.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Kabilan Elangovan and Daniel Ting. When fine-tuning changes the evidence: Architecture-dependent semantic drift. *arXiv preprint arXiv:2604.08513*, 2026.

Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7):620–631, 2021.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *AAAI Conference on Artificial Intelligence*, 2019.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

Arshia Hemmat and Afsaneh Fatemi. Delta-audit: Explaining what changes when models change. *arXiv preprint arXiv:2508.19589*, 2025.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Fabian Hinder, Valerie Vaquet, and Barbara Hammer. Model based explanations of concept drift. *arXiv preprint arXiv:2303.09331*, 2023.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Panagiota Kiourti, Anu Singh, Preeti Duraipandian, Weichao Zhou, and Wenchao Li. Rethinking robustness: A new approach to evaluating feature attribution methods. *arXiv preprint arXiv:2512.06665*, 2025.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

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

Sean Kulinski and David I. Inouye. Towards explaining distribution shifts. In *International Conference on Machine Learning (ICML)*, 2023.

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. Partial order in chaos: Consensus on feature attributions in the Rashomon set. *Journal of Machine Learning Research*, 24(364):1–50, 2023.

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

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011.

Carlos Mougan, Klaus Broelemann, David Masip, Gjergji Kasneci, Thanassis Thiropanis, and Steffen Staab. Explanation shift: How did the distribution shift impact the model? *Transactions on Machine Learning Research (TMLR)*, 2023. arXiv:2303.08081.

Masataka Okabe and Kei Ito. Color universal design: How to make figures and presentations that are friendly to colorblind people. *J*FLY Data Depository for Drosophila Researchers*, 2008.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Shailesh Sridhar, Anton Xue, and Eric Wong. Missingness bias calibration in feature attribution explanations. *arXiv preprint arXiv:2603.04831*, 2026.

Kamalasankari Subramaniakuppusamy and Jugal Gajjar. Feature attribution stability suite: How stable are post-hoc explanations? *arXiv preprint arXiv:2604.02532*, 2026. Accepted at the XAI4CV workshop, CVPR 2026.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

Thackshanaramana B. Hypothesis class determines explanation: Why accurate models disagree. *arXiv preprint arXiv:2603.15821*, 2026.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep Ravikumar. On the (In)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

A TERMINOLOGY: WHAT "EXPLANATION DRIFT" DENOTES

The phrase names at least three distinct objects in work from 2023 onwards, which is why we avoid it. Mougan et al. (2023) use *explanation shift* for a divergence between $P(E_f(X_S))$ and $P(E_f(X_T))$ for a frozen model, a population-level object. Dhayalkar (2026) uses *explanation drift* for the epoch-to-epoch change of attributions on a fixed probe set within a single training run. The concept-drift literature uses *drift explanation* for the opposite direction of inference, explaining why the data moved (Hinder et al., 2023; Kulinski & Inouye, 2023). Our object is none of these: instance-level attribution change across a single discrete model update, measured on prediction-preserved probe points drawn from the shared support. We call it *explanation change* and reserve *drift* for temporal sequences, following the first use.

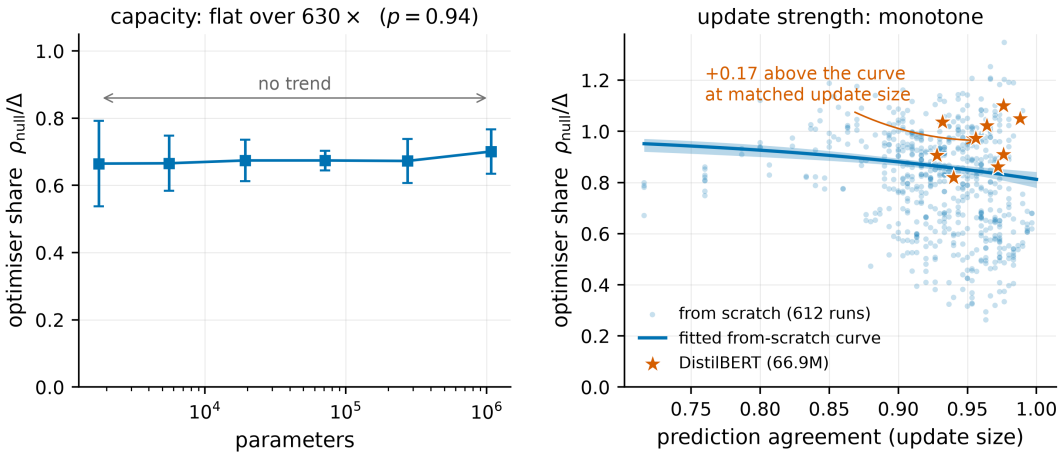


Figure 5: What predicts the share of movement the optimiser reproduces, $\rho_{\text{null}}/\Delta$. Left: capacity does not predict it, with a logit-scale slope of -0.007 $[-0.197, +0.184]$, $p = 0.94$, across a 609× parameter range. Right: update strength does, monotonically; the fitted from-scratch curve is shown with its 95% interval, and DistilBERT sits $+0.173$ $[+0.104, +0.243]$ above it at matched update size.

B THE OPTIMISER SHARE

The share is $\rho_{\text{null}}/\Delta$: the fraction of measured attribution change that the matched null reproduces. A range on its own is a documented correlation; a model of it is a mechanism. Two covariates are comparable across model classes. Prediction agreement measures how far an update moved a model, and capacity measures its size. The width sweep holds the update fixed and varies parameters 609×; the regime sweep holds capacity fixed and varies update strength. The share is bounded, so it is modelled on a logit scale; a linear fit on the raw share extrapolates to impossible negative values.

The share is not bounded by 1 in practice: an update can move the attributions less than its matched null does, and 107 of 681 pooled runs have $\rho_{\text{null}}/\Delta \geq 1$, the largest 1.71. The transform therefore clips to $[0.02, 0.98]$ before taking logits. On the width sweep that carries the capacity fit the clip is almost inert (one of 60 rows is touched and none reaches 1), and the conclusion does not depend on it: dropping that row gives a slope of $+0.073$ $[-0.085, +0.231]$, $p = 0.36$, and fitting the raw share with no transform gives $+0.010$ $[-0.022, +0.043]$, $p = 0.53$, against the reported -0.007 $[-0.197, +0.184]$, $p = 0.94$. All three agree that capacity does not predict the share.

C THE SYNTHETIC GENERATOR

Every synthetic number in this paper comes from `src/uec/data/synthetic.py`. This appendix gives its structural equations, every parameter, the three shift families as parameter changes, and derivations of the two closed forms ω depends on. Within this appendix γ denotes the generator's interaction coefficients, not the integration path of §4, and logit^{-1} is written out so that σ is free for standard deviations.

C.1 BLOCKS

$d = 20$, partitioned into four contiguous blocks.

Block	Indices	Dim.	Role
causal X_C	0–4	5	parents of Y
shortcut X_S	5–7	3	index 5 is a child of Y; 6 and 7 are pure noise
redundant X_R	8–11	4	noisy copies of $X_0, \ldots, X_3$
noise X_N	12–19	8	independent of everything
total		20	

Write $C = \{0, \dots, 4\}$ and $N = \{12, \dots, 19\}$ for the causal and noise index sets.

C.2 Structural equations

In ancestral order, for an environment e with parameters $(\beta, \gamma, \Pi, \mu_c, \sigma_c, \mu_n, \sigma_n, a, \sigma_s, \sigma_r)$:

$$X_C \sim \mathcal{N}\big(\mu_c, \mathrm{diag}(\sigma_c^2)\big), \qquad X_C \in \mathbb{R}^5, \tag{3}$$

$$g(X_C) = \beta^\top X_C + \textstyle\sum_{(j,k)\in\Pi} \gamma_{jk} X_j X_k, \tag{4}$$

$$Y \mid X_C \sim \mathrm{Bernoulli}\big(\mathrm{logit}^{-1}\, g(X_C)\big), \tag{5}$$

$$X_5 \mid Y = a(2Y-1) + \sigma_s \eta_s, \qquad \eta_s \sim \mathcal{N}(0,1), \tag{6}$$

$$X_j \sim \mathcal{N}(0,1), \qquad j \in \{6,7\}, \tag{7}$$

$$X_{8+i} = X_i + \sigma_r \eta_{r,i}, \qquad i = 0,\dots,3,\ \ \eta_{r,i} \sim \mathcal{N}(0,1), \tag{8}$$

$$X_N \sim \mathcal{N}\big(\mu_n, \mathrm{diag}(\sigma_n^2)\big), \qquad X_N \in \mathbb{R}^8, \tag{9}$$

with all η mutually independent and independent of X_C and Y. Coordinates are independent within the causal and noise blocks.

Two structural points carry the rest of the appendix. First, the arrow on X_5 runs from Y, not to it: the shortcut is a child of the label, which is what makes it severable without touching the causal mechanism, and what puts the Bayes logit in closed form. Second, X_R is a measurement of X_C, so it is conditionally independent of Y given X_C while being strongly collinear with it: $\mathrm{Corr}(X_i, X_{8+i}) = (1+\sigma_r^2)^{-1/2} = 0.944$ at $\sigma_r = 0.35$, measured at 0.9438 to 0.9443 over 2×10^5 draws. That is the collinear regime the corollary in §4 requires.

C.3 Parameters

Symbol	Code	Value	Set in
β	`beta`	$(1.2, -0.9, 0.7, 1.0, -0.6)$	`base_environment`
Π	`pairs`	$\{(0,1),(2,3)\}$	`Environment` default
γ	`gamma`	$(0.8, -0.5)$	`base_environment`
μ_c	`mu_c`	$\mathbf{0}_5$	`Environment` default
σ_c	`sd_c`	$\mathbf{1}_5$	`Environment` default
μ_n	`mu_n`	$\mathbf{0}_8$	`Environment` default
σ_n	`sd_n`	$\mathbf{1}_8$	`Environment` default
a	`a`	0; 0.8 in the shortcut source	`make_pair`, via `shortcut_a`; never overridden by a caller
σ_s	`sigma_s`	0.8	`Environment` default, never overridden
σ_r	`sigma_r`	0.35	`Environment` default, never overridden
m	`magnitude`	1.5	every runner; the regime sweep uses $\{0.5, 0.75, 1.25, 2.0\}$
b	`BASELINE`	$\mathbf{0}_{20}$	`run_synthetic.py`; ablated against the source and target means
τ	`tau`	2.0	`shared_support_probe`

Sample sizes are 8000 source, 4000 update and 500 probe, the last cut to 100 for KernelSHAP and LIME, which are quadratic in cost. They live in `configs/synthetic.yaml` and are held to the runner's own argparse defaults by `tests/test_configs.py`, so the config cannot drift from what was run. `make_pair`'s own default magnitude is 0.75; every runner passes 1.5, so 0.75 is never used.

C.4 The three shift families

A shift family is a parameter change and nothing else. `make_pair` builds the source, then copies it with the listed fields replaced; every field not listed is carried over identically.

Family	Knob	Source → target
`none` (placebo)	—	target is a field-for-field copy
`covariate`	μ_c	$\mathbf{0}_5 \to m\mathbf{1}_5 = 1.5 \cdot \mathbf{1}_5$
	σ_c	$\mathbf{1}_5 \to (1 + 0.3m)\mathbf{1}_5 = 1.45 \cdot \mathbf{1}_5$
	μ_n	$\mathbf{0}_8 \to 0.5m\mathbf{1}_8 = 0.75 \cdot \mathbf{1}_8$
`concept`	β	$(1.2, -0.9, 0.7, 1.0, -0.6) \to (1.2, 0.9, -0.7, 1.0, 0.8)$ entries 1 and 2 sign-flipped, entry 4 shifted by $+1.4$
	γ	$(0.8, -0.5) \to (0.2, 0.5)$, elementwise by $(0.25, -1.0)$
`shortcut`	a	$0.8 \to 0$

The covariate family moves $P(X)$ only, on the causal and noise blocks; the concept family moves $P(Y \mid X_C)$ only; shortcut removal severs X_5 from Y and leaves the causal mechanism and $P(X_C)$ untouched. The magnitude m enters the covariate family alone, though the runners pass it uniformly. Marginal label rates over 10^5 draws are 0.523 at the source, 0.753 at the covariate target ($P(Y \mid X)$ is fixed; $P(Y)$ moves because $P(X)$ did), 0.503 at the concept target and 0.523 at the shortcut target. Since $a = 0$ in the `none`, `covariate` and `concept` families, coordinate 5 is $\mathcal{N}(0, \sigma_s^2)$ noise there: the shortcut exists in the shortcut family only.

C.5 The Bayes-optimal logit, derived

Equations (3)–(9) factorise the joint as

$$p_e(x, y) = p(x_C)\, P(y \mid x_C)\, p(x_5 \mid y)\, p(x_6)p(x_7)\, p(x_R \mid x_C)\, p(x_N). \tag{10}$$

Exactly two factors depend on y. Taking the ratio at $y = 1$ and $y = 0$, every other factor cancels, so

$$h_e(x) = \operatorname{logit} P_e(Y{=}1 \mid X{=}x) = \log \frac{P(1 \mid x_C)}{P(0 \mid x_C)} + \log \frac{p(x_5 \mid 1)}{p(x_5 \mid 0)}. \tag{11}$$

The first term is $g(x_C)$ by (5). The second is a Gaussian likelihood ratio with common variance, so the quadratic terms cancel:

$$\log \frac{p(x_5 \mid 1)}{p(x_5 \mid 0)} = \frac{-(x_5 - a)^2 + (x_5 + a)^2}{2\sigma_s^2} = \frac{2a}{\sigma_s^2}\, x_5. \tag{12}$$

Writing $c_e = 2a_e/\sigma_{s,e}^2$,

$$\boxed{h_e(x) = \beta^\top x_C + \sum_{(j,k)\in\Pi} \gamma_{jk} x_j x_k + c_e\, x_5} \tag{13}$$

which is `Environment.bayes_logodds`.

The closed form is available, rather than merely asserted, for three reasons visible in the derivation. The redundant block, the unused shortcut coordinates and the noise block cancel because each is conditionally independent of Y given X_C: X_R is informative about X_C, but X_C is observed, so it adds nothing. The label model (5) is logistic by construction, so its contribution is $g(x_C)$ with no normaliser. And the shortcut is a Gaussian child of Y with label-independent variance, so its likelihood ratio is linear in x_5. Change any of the three and the posterior stops being elementary; keep them and h_e is a quadratic polynomial supported on the five causal coordinates plus coordinate 5.

Two consequences are used later. $c = 2a/\sigma_s^2 = 2.5$ at $a = 0.8, \sigma_s = 0.8$, and $c = 0$ whenever $a = 0$, so the reference is supported on exactly five coordinates in the `none`, `covariate` and `concept` families and on six in the shortcut source. And (13) is a calibration claim, not a definition, so it is testable: `tests/test_synthetic.py::test_bayes_logodds_matches_empirical` draws 4×10^5 samples, bins them into twenty quantile bins of h, and requires the empirical label rate in each bin to match $\operatorname{logit}^{-1}(\bar{h})$ to within 0.025, for both environments of all four families.

C.6 INTEGRATED GRADIENTS OF THAT LOGIT, DERIVED

For a function h, baseline b and straight path $b + \alpha(x - b)$, $\mathrm{IG}_j[h](x, b) = (x_j - b_j) \int_0^1 \partial_j h(b + \alpha(x - b))\, \mathrm{d}\alpha$. From (13), writing $k \sim j$ when $(j, k) \in \Pi$ or $(k, j) \in \Pi$ and γ_{jk} for that pair's coefficient,

$$\partial_j h(z) = \beta_j + \sum_{k \sim j} \gamma_{jk} z_k \quad (j \in C), \qquad \partial_5 h(z) = c, \qquad \partial_j h(z) = 0 \text{ otherwise}, \tag{14}$$

so the gradient is affine in z and z is affine in α. The integrand is therefore affine in α and the integral is its midpoint value, using $\int_0^1 (b_k + \alpha(x_k - b_k))\, \mathrm{d}\alpha = (b_k + x_k)/2$:

$$\mathrm{IG}_j[h](x, b) = \begin{cases} (x_j - b_j)\Big(\beta_j + \sum_{k \sim j} \gamma_{jk} \dfrac{b_k + x_k}{2}\Big), & j \in C, \\ c\,(x_5 - b_5), & j = 5, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

which is `groundtruth.gt_attributions`. Nothing in the derivation used the input distribution, so (15) is exact for every x and every baseline b, including the source-mean and target-mean baselines used in the ablation. Completeness holds identically rather than in the limit: the two halves of an interaction sum to $\frac{\gamma_{jk}}{2}\big[(x_j - b_j)(b_k + x_k) + (x_k - b_k)(b_j + x_j)\big] = \gamma_{jk}(x_j x_k - b_j b_k)$, so $\sum_j \mathrm{IG}_j = h(x) - h(b)$.

Numerical confirmation and where it is established. The closed form is checked against 512-step midpoint quadrature of the same integral, and separately against completeness.

Setting	Rows	Quadrature gap	Completeness residual
generator, 8 environments	2000 each	3.6×10^{-14}	3.6×10^{-15}
semi-synthetic, real ACS rows	2000	1.1×10^{-14}	1.8×10^{-15}

The quadrature gap is $\max_{x,j} |\mathrm{IG}_j - \widehat{\mathrm{IG}}_j|$ against the 512-step estimate and the completeness residual is $\max_x |\sum_j \mathrm{IG}_j - (h(x) - h(b))|$. The second row is the 1.1×10^{-14} quoted in §6. It is produced by `experiments/run_semisynthetic.py`, in the block headed "1. Is the closed form correct on real rows?", which compares `Mechanism.ig` against `Mechanism.ig_quadrature` on 2000 real ACS rows, asserts agreement below 10^{-8} before any other result is computed, and writes the value to the `quadrature_err` column of `T18_semisynthetic.csv`. The generator's counterpart is `groundtruth.gt_attributions_quadrature`, pinned at 10^{-9} by `tests/test_synthetic.py::test_ig_closed_form_matches_quadrature` and at 10^{-9} for completeness by `::test_ig_completeness`, in both environments of all four families. Both residuals sit at the double-precision floor for the magnitudes involved, so the quadrature is the approximation and (15) is the exact object.

C.7 WHY ω VANISHES UNDER COVARIATE SHIFT

Proposition 4 (Exact null under covariate shift)**.** *In the covariate family, $\omega(x) = 0$ for every $x \in \mathbb{R}^{20}$, every baseline b, every normaliser ϕ and every distance d with $d(u, u) = 0$.*

Proof. The covariate family replaces (μ_c, σ_c, μ_n) and nothing else, so $(\beta, \gamma, \Pi, a, \sigma_s)$ are carried to the target unchanged. By (13) the Bayes logit depends on the environment only through $(\beta, \gamma, \Pi, a, \sigma_s)$: the parameters of $P(X_C)$ and $P(X_N)$ entered (10) as factors common to the $y = 1$ and $y = 0$ cases and cancelled. Hence $h_S \equiv h_T$ as functions on $\mathbb{R}^{20}$. Integrated gradients are a functional of h, x and b alone, so $\mathrm{IG}[h_S](x, b) = \mathrm{IG}[h_T](x, b)$ pointwise, and $\omega(x) = d\big(\phi(A), \phi(A)\big) = 0$. □

The vanishing is exact and not merely small. Source and target attributions are evaluated by the same code from bitwise-identical parameter arrays, so the two arrays are bitwise equal; `tests/test_synthetic.py::test_omega_is_zero_under_covariate_shift_`

`and_positive_otherwise` asserts `np.array_equal` on the raw attributions and $\omega < 10^{-12}$ under all four distances, and asserts $\omega > 0$ under concept and shortcut shift. Over 4000 samples the primary pair (ℓ_1-normalised magnitudes, ℓ_1 distance) gives mean $\omega = 0.115$ under concept shift and 0.370 under shortcut removal, against exactly 0 under covariate shift and the placebo.

Under the other two families the difference is explicit. Concept shift moves β by $(0, 1.8, -1.4, 0, 1.4)$ and γ by $(-0.6, 1.0)$, so (15) changes on all five causal coordinates. Shortcut removal moves c from 2.5 to 0, so $\mathrm{IG}[h_T] - \mathrm{IG}[h_S]$ is supported on coordinate 5 alone before normalisation and ϕ spreads it over the vector.

One distance is blind here by construction and is retained only as an ablation axis. With $a = 0$ the reference is supported on exactly five coordinates, so top-5 Jaccard cannot register a concept shift that only reweights those five, and returns $\omega = 0$; top-3 registers it. The primary distance is ℓ_1, and the blind case is pinned by `::test_topk5_is_blind_to_concept_shift_by_construction`.

C.8 Shared support

The log density is closed form as well, which makes probe membership exact rather than estimated:

$$\begin{aligned}\log p_e(x) = &\sum_{j\in C} \log \mathcal{N}(x_j; \mu_{c,j}, \sigma^2_{c,j}) + \sum_{j\in N} \log \mathcal{N}(x_j; \mu_{n,j}, \sigma^2_{n,j}) \\ &+ \sum_{i=0}^{3} \log \mathcal{N}(x_{8+i}; x_i, \sigma_r^2) + \sum_{j\in\{6,7\}} \log \mathcal{N}(x_j; 0, 1) \\ &+ \log\Big[\pi(x_C)\,\mathcal{N}(x_5; a, \sigma_s^2) + (1-\pi(x_C))\,\mathcal{N}(x_5; -a, \sigma_s^2)\Big], \end{aligned} \tag{16}$$

with $\pi(x_C) = \mathrm{logit}^{-1}\, g(x_C)$. The last line is a two-component mixture because X_5 is a child of the latent Y; the rest is a product of Gaussians. The probe draws a balanced source/target mixture, keeps the rows with $|\log p_T(x) - \log p_S(x)| \le \tau = 2$ and subsamples to n_{probe} without replacement, so neither checkpoint extrapolates on it.

Overlap under the source, $\Pr_{X\sim P_S}\big(|\log(p_T/p_S)| \le 2\big)$, falls with the shift magnitude and bounds what can be studied:

magnitude m	0.5	0.75	1.0	1.25	1.5	2.0
overlap	0.83	0.54	0.32	0.18	0.10	0.027

(mean of five draws of 2×10^4 source samples each). The headline setting $m = 1.5$ leaves 10% overlap; beyond $m \approx 2$ there is essentially no region in which both checkpoints are in distribution, so instance-level comparison stops being meaningful. The identity $\mathbb{E}_{X\sim P_S}[p_T(X)/p_S(X)] = 1$ validates both densities at once and is pinned to 5% over 3×10^5 samples in all four families by `::test_density_ratio_importance_identity`.

D Architectures, training and explainer configuration

Every value below is a default in the released code or a recorded field in `results/registry.csv`. Where a runner default and a configuration file could drift, `tests/test_configs.py` asserts they agree.

D.1 Model arms

Initialisation. MLPs and the CIFAR ResNet use the PyTorch defaults for `nn.Linear` and `nn.Conv2d` (Kaiming-uniform weights on fan-in, uniform biases), drawn after `seed_everything(seed)`. DistilBERT starts from the `distilbert-base-uncased` checkpoint with a randomly initialised two-way classification head. The tree arm has no initialisation; `random_state` seeds the subsampling.

Table 2: One row per model arm. Layer sizes are written input–hidden–output. Parameter counts for the MLPs are weight counts, the convention the released artefacts record; including biases adds $2w + 1$ (Table 3). Update columns give n / learning rate / epochs / batch. The synthetic MLP and the width sweep share an operator; the tree arm has no fine-tune and retrains instead.

Arm	Architecture	Params	Act.	Optimiser	Source train	Update
Synthetic MLP	20–64–64–1	5,440	SiLU	Adam, λ=10^{-5}	8,000 / 2×10^{-3} / 60 / 256	4,000 / 2×10^{-4} / 2 / 256
Width sweep	20–w–w–1	1.8k–1.07M	SiLU	Adam, λ=10^{-5}	8,000 / 2×10^{-3} / 60 / 256	4,000 / 2×10^{-4} / 2 / 256
XGBoost	200 trees, depth 4	—	—	boosting, λ_2=1	8,000 / 0.1 / 200 trees	4,000 / 0.1 / 200 trees
CIFAR ResNet	stem + 3 blocks, 16–32–64	78,042	SiLU	Adam, λ=0	10,000 / 2×10^{-3} / 15 / 128	14,000 / 2×10^{-4} / 1 / 128
DistilBERT	6 layers, d=768	66,955,010	GELU	AdamW, λ=0.01	5,000 / 2×10^{-5} / 1 / 16	4,000 / 10^{-6}–10^{-5} / 1 / 16

Losses and devices. MLPs use `BCEWithLogitsLoss` on a single logit; the ResNet and DistilBERT use cross-entropy. The synthetic, tree, ACS and CIFAR arms run on CPU with the thread count pinned, since CPU reductions are bitwise reproducible only at a fixed thread count. DistilBERT requires one GPU; its output is committed so the rest reproduces without it.

Update semantics. On the synthetic, ACS and semi-synthetic arms the update is a fine-tune on n_{update} fresh points. On CIFAR and DistilBERT it is *additive*: the new data is concatenated with a replay slice of the source, identical in both arms, and the pair is trained together for one epoch. Replay is the whole 10,000-image source on CIFAR and a 2,000-review slice on DistilBERT. Without replay the update is domain replacement, which caused catastrophic forgetting on CIFAR (accuracy $0.59 \to 0.44$, agreement 0.54) and destroyed the prediction-preserved probe the comparison rests on.

Explained scalar. The MLPs emit one logit. The ResNet and DistilBERT are wrapped so they emit the logit of a *fixed* class, the source model's own arg-max, held constant across checkpoints; explaining each checkpoint's arg-max would difference attributions for two different quantities.

D.2 Activation, and why ReLU is only an ablation

The code default is `silu`; `relu`, `gelu` and `softplus` are available and only `silu` is used for any reported MLP or CIFAR result. Proposition 1 is stated for C^1 functions. A ReLU network is not one: its gradient is piecewise constant, so Riemann quadrature for IG converges as $O(1/n_{\text{steps}})$ and completeness holds to about 2×10^{-3} at 512 steps. That residual is the same order as the quantity §6.6 tests, so a ReLU default would let numerical error masquerade as a bound violation. ReLU is therefore carried as an ablation, where the ordering of conclusions is unchanged. DistilBERT's GELU and the tree arm's piecewise-constant response are not ours to choose; neither enters the Proposition 1 check.

D.3 Width ladder

The quoted $609\times$ range is $1{,}072{,}129/1{,}761 = 608.8$, counting every weight and bias: $w^2 + 23w + 1$ for the 20–w–w–1 stack. Each rung uses covariate shift at magnitude 1.5, n_{source}=8,000, n_{update}=4,000, a 300-point shared-support probe, 5 seeds, and integrated gradients and gradient×input only. IG runs at 32 steps rather than 64 on this arm; both checkpoints use the same quadrature, so the ratio is unaffected and completeness still holds to $\sim 4 \times 10^{-4}$. The sweep builds the matched null and the treatment only, so no seed floor is available for it.

Table 3: The width sweep. Weight counts are what the artefacts record and what the ratio quoted in the main text uses; the all-parameter column adds the three bias vectors. The sweep holds the operator fixed at the synthetic headline setting and varies only w.

Hidden width w	32	64	128	256	512	1024
Weights ($20w + w^2 + w$)	1,696	5,440	19,072	70,912	272,896	1,070,080
All parameters	1,761	5,569	19,329	71,425	273,921	1,072,129

D.4 Explainer hyperparameters

Table 4: Explainer settings. All values are the function defaults in `src/uec/explain/` and are mirrored in `configs/explainers.yaml`. The gradient family is evaluated in float64: Proposition 1(i) is an equality tested per probe point, and in float32 Captum's completeness error is $\sim 5 \times 10^{-3}$ on logits of order 5, the same size as the quantity under test. Seeds are written as functions of the repetition index r, which indexes the two runs used for the noise floor ν.

Explainer	**Family**	**Draw**	**Settings**
Saliency	gradient	exact	—
Gradient×input	gradient	exact	—
Integrated gradients	path	exact	64 steps, midpoint Riemann, baseline $b = \mathbf{0}$
Expected gradients	path	stochastic	32 samples, added noise 0, background 25, seed $9000{+}17r$
SmoothGrad	gradient	stochastic	32 samples, $\sigma = 0.15\times$ per-feature s.d. of the probe, seed $4000{+}17r$
KernelSHAP	Shapley	stochastic	128 coalition samples, background 25, `l1_reg=num_features(20)`, seed $7000{+}31r$
LIME	surrogate	stochastic	1,500 samples, regression mode on the logit, no discretisation, library default kernel width, seed $5000{+}31r$
TreeSHAP	Shapley	exact	XGBoost native; tree arm only, so $\nu = 0$ by construction

Baseline and probe. The attribution baseline is $b = \mathbf{0}$ in the standardised input space, fixed across every checkpoint and every environment, and it is the same baseline ω integrates from, which is what makes the two comparable. The background set for expected gradients, KernelSHAP and LIME is 25 points drawn from the *source* environment per (seed, family). The five cheap explainers run on all 500 probe points; KernelSHAP and LIME are quadratic in cost and run on the first 100, a nested prefix, so every explainer is compared on a common subset. Attributions are cached by a key containing the checkpoint weight hash, so editing a model cannot serve a stale array.

D.5 The matched-operator null, operationally

Four checkpoints are built per (seed, shift family) by `build_checkpoints` in `uec.train.harness`. Each draws from its own named stream, `default_rng([seed, k])`, rather than from a shared sequential generator, so a caller that requests a subset of regimes does not silently shift the draws of the rest.

- **Source** (k=0): train from initialisation on 8,000 source points, 60 epochs, 1,920 steps, lr 2×10^{-3}.
- **Matched null** (k=1): fine-tune the source checkpoint on 4,000 *fresh source* points, 2 epochs, 32 steps, lr 2×10^{-4}, update seed seed+1.
- **Treatment** (k=2): the identical call on 4,000 *target* points, same epochs, same step count, same learning rate, same update seed seed+1.
- **Seed floor**: re-train from a fresh initialisation, seed seed+5000, on the source model's *own* 8,000 training rows (60 epochs, 1,920 steps, lr 2×10^{-3}). Resampling the data as well

would confound seed variation with sampling variation, so this is the floor as Bensmail (2025) and Laberge et al. (2023) measure it.

The contract is enforced, not asserted in prose. `operator_signature(meta)` returns $(n_{\text{train}}, n_{\text{steps}}, \text{lr}, \text{epochs})$ and the equality of the null's and the treatment's signatures is an `assert` in every runner: `run_synthetic.py`:80, `models/trees.py`:63, `run_folktables.py`:64, `run_semisynthetic.py`:79, and a step-count equality in `run_vision.py`:129 and `kaggle/scale_probe.py`. Any asymmetry in those four quantities aborts the run.

What the seed floor does not match. Against the update, the seed floor differs in the operator (from-scratch versus fine-tune), the data volume (8,000 versus 4,000), the learning rate (2×10^{-3} versus 2×10^{-4}) and the step count (1,920 versus 32, a factor of 60). This is the arithmetic behind §6.2: a fixed quantity that large is not a floor for a 32-step update, and the ratio to it tracks how hard one trained.

Where the operator differs by arm. The tree arm has no fine-tune, so source, null, treatment and seed floor are four full retrains with identical hyperparameters; this makes the matched null exact rather than approximate. The ACS arm trains the source for 40 epochs on 20,000 rows and updates at 2 and 20 epochs. The semi-synthetic arm updates for 20 epochs at lr 5×10^{-4} (roughly ten times the step count of the synthetic headline), which is part of why its completeness is 1.09 rather than 0.12 or 0.54. The CIFAR and DistilBERT arms have no from-scratch-with-a-different-seed floor built on matched data; DistilBERT reports ρ_{seed} equal to ρ_{null}, so no seed-floor claim is made there.

D.6 Seeds and testing

Ten seeds for the synthetic, tree, ACS and semi-synthetic arms; five for the regime sweep (600 runs: 4 magnitudes × 3 update learning rates × 5 epoch settings × 5 seeds × 2 explainers) and for the width sweep; three for CIFAR and DistilBERT. `seed_everything` fixes the Python, NumPy and Torch generators and requests deterministic algorithms. Each stochastic explainer is run twice on every checkpoint, at $r = 0$ and $r = 1$, and the distance between the two runs is the noise floor ν; deterministic explainers reproduce bitwise, so $\nu = 0$ exactly. Intervals are bootstrapped over seeds, not over probe points. Paired tests compare the null and the treatment within seed, with Holm correction and Cliff's δ.

E How each shift is instantiated

E.1 CIFAR-10 corruptions

Corruptions are generated locally, not downloaded: the CIFAR-10-C archive is 2.9 GB and the families used here are a few lines each. They serve only as a covariate-shift *generator* for the model update. Two are applied, in this order, to the new half of the additive update:

- **Gaussian noise**, $X \to \text{clip}(X + \varepsilon,\, 0,\, 1)$ with $\varepsilon \sim \mathcal{N}(0, 0.08^2)$ per pixel and channel, on $[0, 1]$ images.
- **Contrast**, $X \to \text{clip}\big((X - m)\, 0.4 + m,\, 0,\, 1\big)$ with m the per-image, per-channel mean.

Brightness ($+0.25$) and pixelate (block size 3) are defined in the same module and are not invoked by any reported run; the pixelate severity does not divide 32 and would raise on CIFAR input. The spurious variant plants a class-parity tint of ± 0.35 on the red channel of a 6×6 corner patch, present in the source and in the probe, removed from the new half. Probes are drawn from the clean test set so both checkpoints are in distribution on them.

Both CIFAR families are reported with ω entered as a zero vector. For the corruption family that is the covariate-shift null. For the shortcut family it is an understatement we state rather than hide: a model that stops using a removed tint *should* move its attributions, so the true ω there is positive and the reported unwarranted change is an upper bound that we do not correct. No warranted-change claim is made on CIFAR.

Table 5: Shift families by dataset. "Exact" means the Bayes-optimal predictor of both environments is known in closed form, so ω is computed rather than estimated. "Entered as 0" means the arm reports Δ against ρ_{null} with a zero reference supplied in code: no warranted-change claim is made there.

Dataset	Family	Instantiation	ω
Synthetic SEM	none	target environment is the source environment	exact, $= 0$
	covariate	$\mu_C \to \mu_C + 1.5$, $\sigma_C \to 1.45\,\sigma_C$, $\mu_N \to \mu_N + 0.75$; $P(Y \mid X)$ untouched	exact, $= 0$
	concept	β_1, β_2 sign-flipped, $\beta_4 \to \beta_4 + 1.4$, $\gamma \to (0.25\gamma_1, -\gamma_2)$	exact, closed form
	shortcut	shortcut coupling $a : 0.8 \to 0$, decoupling S_0 from Y	exact, closed form
CIFAR-10	corruption	additive update; the new half is corrupted locally, Gaussian noise then contrast	entered as 0
	shortcut	corner tint planted in the source and in the probe, absent from the new half	entered as 0
ACS Income	state	CA 2018 $\to$ MI, MS, SD, PR; source-fitted scaler frozen	not available; checked
	year	CA 2018 $\to$ CA 2022 on the nine cross-year features	not available; checked
ACS semi-syn.	covariate	real rows resampled $\propto \exp\big(2\,\text{clip}(z_{\text{AGEP}}, -3, 3)\big)$	exact, $= 0$
	concept	$\beta_0 + 0.8$, $\beta_3 - 0.6$, $\beta_6 + 0.5$ on the fitted mechanism	exact, closed form
IMDB	domain	additive update; the new half is Rotten Tomatoes	entered as 0

E.2 The semi-synthetic ACS construction

The generator supplies both the covariates and the mechanism, which leaves the objection that the closed form might only behave on well-conditioned Gaussians. This construction removes the covariate half of that assumption: the rows are real ACS records and only the labels are regenerated.

Covariates. ACS Income, California 2018: 195,665 rows and the ten `ACSIncome` features, standardised by a scaler fitted on the source and frozen. A per-checkpoint or per-domain scaler would change the input parameterisation between checkpoints and make attribution differences uninterpretable.

Mechanism. A quadratic logit on the standardised covariates,

$$h(x) = \beta^\top x + \textstyle\sum_{(j,k)\in\mathcal{I}} \gamma_{jk}\, x_j x_k, \qquad \mathcal{I} = \{(0,2),(3,7)\},\ \gamma = (0.4,\ -0.3), \tag{17}$$

with β taken from an ℓ_2-regularised logistic regression (C=1, 2,000 iterations) fitted to the *real* income labels. Fitting β to the real outcome keeps the labelling task a plausible one for these features; an arbitrary β would make the covariates uninformative in an unrealistic way. The interactions are added so the reference is not linear, since a linear logit makes IG trivially $\beta_j(x_j - b_j)$ and would not exercise the closed form. The interacting pairs are age $\times$ education and marital status $\times$ hours worked. Labels are then drawn as $Y \sim \text{Bernoulli}(\sigma(h(x)))$ from a seed-indexed generator, which makes h Bayes-optimal by construction.

Closed form and its check. The path integral collapses analytically for any x:

$$\text{IG}_j(x; b) = (x_j - b_j)\Big(\beta_j + \textstyle\sum_k \gamma_{jk}\,(b_k + x_k)/2\Big). \tag{18}$$

This is a property of the polynomial, not of the data. Re-deriving it by 512-step midpoint quadrature on 2,000 real rows gives a maximum absolute disagreement of 1.11×10^{-14} and a maximum completeness residual of 1.8×10^{-15}, so the machinery transfers to real covariates exactly.

The tilt. Covariate shift is a reweighting, never an edit: rows are real and unmodified and only their frequency changes, so the mechanism is untouched and ω is exactly zero. Target rows are drawn with replacement with weight $\exp(2.0\,\mathrm{clip}(z_{\text{AGEP}}, -3, 3))$. Both arms draw with replacement; drawing the null without it would hand the treatment fewer distinct rows and the ratio would partly measure effective sample size, which is the confound this paper is about.

Calibration of the tilt strength. The strength is set by a cross-fitted domain classifier, not chosen. At strength 2.0 a four-fold cross-fitted logistic classifier separates the tilted update sample from an untilted draw of the same size at AUC 0.901 (five seeds, range 0.896–0.905); at strength 1.0 it reaches only 0.758, so a null result there would have measured a weaker shift rather than a weaker effect. The synthetic covariate generator reaches the same AUC, 0.902, at magnitude 0.75. It is worth being explicit that the headline synthetic results use magnitude 1.5, whose domain AUC is 0.992: the tilt is matched to a milder synthetic shift than the one §6.3 reports, which is consistent with its smaller ratio (1.20 against 1.52) and means the semi-synthetic arm should be read as a transfer check on the machinery rather than as a replication of the headline magnitude.

Concept shift. The mechanism itself is perturbed, $\beta_0{+}0.8$, $\beta_3{-}0.6$, $\beta_6{+}0.5$: two features the interactions touch and one they do not, so the warranted change is not concentrated in a single coordinate. ω is then non-zero and still closed-form.

E.3 ACS Income: real shifts, and what stands in for ω

The state and year shifts are observed, not constructed, so no Bayes predictor is available and ω is not computed. Two diagnostics stand in. A cross-fitted domain classifier estimates the log density ratio, since its logit is that ratio up to the class-prior offset, and points are kept only where the estimate lies within $\tau = 0.75$ of the prior-corrected origin. Pooled domain AUC is 0.71 (MI), 0.72 (MS), 0.71 (SD), 0.61 (PR) and 0.64 (CA 2022); within the retained probe it falls to 0.59–0.63, which is what a genuine overlap looks like. Update size is set to what the *smaller* domain can supply and used by both arms, so a small target state cannot silently give the null more data than the treatment.

The second diagnostic, `mechanism_stability`, is the reliability-curve distance between the source model's calibration on source and on target points within the shared support. It is a necessary condition for $P(Y \mid X)$ stability, not a sufficient one. It reads 0.064 for MI, 0.087 for CA 2022, 0.142 for SD, 0.203 for MS and 0.287 for PR. The two large values are worth naming: on MS and PR the covariate-shift null is the weakest, and the corresponding cells should be read as the least secure in the ACS set.

E.4 Is the covariate-shift assumption tenable here?

Under covariate shift $P(X)$ moves and $P(Y \mid X)$ does not, so the Bayes-optimal predictor is unchanged and $\omega = 0$ exactly. That is a construction, not an inference, on the synthetic generator and on the semi-synthetic tilt: in the first the conditional is written down, in the second the labelling mechanism is held fixed while only the sampling weights change. Nothing needs defending in those two settings.

CIFAR is the setting where the assumption is an assumption. The corruptions used here are mild: additive noise at $\sigma = 0.08$ on $[0, 1]$ pixels and a contrast multiplier of 0.4. At that severity the class remains the class (the images are noisier and flatter, not ambiguous), so treating the optimal predictor as unchanged is defensible, and the arm reports $\omega = 0$ on that basis. The ResNet itself is weak in absolute terms, at 0.596 source accuracy under corruption and 0.689 under the planted shortcut: it is a sanity check that the effect survives a convolutional architecture, not a vision benchmark, and we do not read its ratios as competitive numbers.

The honest form of the argument is that mildness is what licenses $\omega = 0$, not that the assumption is robust. At CIFAR-10-C severity 5 it would fail, and fail in a way that matters: heavy corruption destroys class evidence rather than merely perturbing it, so the Bayes-optimal conditional genuinely moves, ω becomes non-zero and unknown, and $\Delta - \omega$ stops being interpretable. A measurement that continued to report $\omega = 0$ there would charge genuine adaptation as unwarranted change, the precise error §3 criticises in the prior work. We therefore do not claim the effect is robust to corruption severity. We claim the opposite: the design holds severity low so that the reference is

defensible, and the unwarranted change it finds cannot be explained away as a shifted conditional. The same reasoning is why DistilBERT's IMDB → Rotten Tomatoes shift is reported with ω entered as zero and no warranted-change claim: both are film reviews with the same label semantics, which makes the null plausible, but nothing here establishes it.

F THE TWO REIMPLEMENTED METRICS, AND THE RE-AUDIT'S RESAMPLING

Both competitor metrics are recomputed on our own checkpoints, probe and attributions, so the comparison in §6.4 is computed rather than asserted. This appendix states each as an algorithm, names the cell the quoted percentages come from, and fixes the resampling scheme the re-audit uses.

F.1 SHARED MACHINERY

Write $A_t, A_{t+1} \in \mathbb{R}^{n\times d}$ for the probe attributions at the two checkpoints. The normaliser is attribution mass, row-wise,

$$\phi(A)_j = |A_j| \Big/ \textstyle\sum_k |A_k|, \tag{19}$$

returning zeros on an all-zero row (`normalise.py:8--11`). The distance is half the L_1 gap, $d(u,v) = \frac{1}{2}\sum_j |u_j - v_j| \in [0,1]$ (`distances.py:46--47`), which is the primary choice (`distances.py:63`). The preserved set is $\mathcal{P}_\epsilon$ with $\epsilon = 0.05$ applied to probabilities (`uec.py:19--20`). The run is 5 seeds, shift magnitude 1.5, update learning rate 5×10^{-4}, 8,000 source and 4,000 update points, a 400-point probe, and update budgets of 2, 20, 100 and 400 epochs. Both metrics below consume exactly these objects, on exactly these points.

F.2 THE FASS-STYLE FILTERED DISTANCE

Subramaniakuppusamy & Gajjar (2026) condition on prediction preservation and then report the raw distance. As implemented (`baselines.py:22--25`, invoked at `baselines.py:60`):

1. Compute $\Delta(x) = d\big(\phi(A_t(x)), \phi(A_{t+1}(x))\big)$ at every probe point (`run_differentiation.py:78`).
2. Restrict to $\mathcal{P}_\epsilon$.
3. Report $\mathbb{E}_{\mathcal{P}_\epsilon}[\Delta]$, or NaN if that subset is empty.

No reference and no floor are subtracted. The metric is therefore numerically identical to our $\mathbb{E}_{\mathcal{P}_\epsilon}[\Delta]$, and the two columns agree to machine precision on all 180 rows of the artefact. Everything that separates it from UEC is the two subtracted terms.

F.3 DELTA-AUDIT'S SPURIOUS RESIDUAL

Hemmat & Fatemi (2025) call the part of attribution movement that behaviour change does not explain "risky reliance redistribution". As implemented (`baselines.py:28--40`):

1. Restrict to $\mathcal{P}_\epsilon$; every step below runs on that subset alone (`baselines.py:55`).
2. Attribution movement, *without* the factor of $\frac{1}{2}$, so $m_i = \sum_j \big|\phi(A_t)_{ij} - \phi(A_{t+1})_{ij}\big| = 2\Delta(x_i)$ (`baselines.py:34`).
3. Behaviour movement $b_i = |p_t(x_i) - p_{t+1}(x_i)|$, on probabilities and not on logits (`baselines.py:35`).
4. Fit a degree-one least-squares line, $(\hat{a}, \hat{c}) = \arg\min_{a,c} \sum_i (m_i - ab_i - c)^2$ (`baselines.py:39`).
5. Report $\mathbb{E}\big[\,|m_i - \hat{a}b_i - \hat{c}|\,\big]$, or NaN if either m or b has standard deviation below 10^{-12} (`baselines.py:36--37`).

Two consequences follow from step 1, which is our choice and not theirs, made so that both metrics see the points UEC sees. The regressor b_i is confined to $[0, \epsilon]$ by construction, so the line is fitted

inside a window of width 0.05; and the residual is scored against a line, so any non-linear dependence on behaviour change is booked as spurious.

F.4 Where 45% and 89% come from

Both percentages come from a single cell of `T4_differentiation.csv`: explainer integrated gradients, update budget 100 epochs, shortcut removal against covariate shift, averaged over the 5 seeds. Under covariate shift ω is exactly zero, so all measured change is drift; under shortcut removal the Bayes-optimal predictor genuinely stops using S_0, so most of it is adaptation the model was right to make.

	covariate shift	shortcut removal	shortcut / covariate
FASS filtered distance	0.19230	0.27860	$+44.9\%$
Delta-Audit spurious residual	0.08004	0.15121	$+88.9\%$
warranted reference ω	0.00000	0.24652	
matched null ρ_{null}	0.17716	0.17529	
UEC	$+0.01514$	-0.14320	

The arithmetic is $0.27860/0.19230 = 1.4488$ and $0.15121/0.08004 = 1.8891$, that is 45% and 89% higher on the case that should score better. Pairing within seed before averaging gives 45.3% and 89.0%, so the quantity does not depend on the order of the two operations.

This is one of twelve explainer $\times$ budget cells, and we state the rest rather than leave it to be found. The FASS gap is positive in all twelve, running from $+3.3\%$ to $+91.8\%$. The spurious-residual gap runs from -39.0% to $+133.1\%$ and is negative in all three 2-epoch cells, where neither model has moved far enough for the shortcut to be dropped. The quoted cell is representative of the budgets at and above 20 epochs and not of the lightest one.

F.5 The opposite signs

UEC applies Eq. (2) to the same two runs. Under covariate shift $0.19230 - 0 - 0.17716 = +0.01514$; under shortcut removal $0.27860 - 0.24652 - 0.17529 = -0.14320$. Positive means movement the floors and the reference cannot account for, negative means the update moved less than warranted change plus the optimiser alone would predict. The two shift families are therefore separated in sign, not merely in magnitude, which is the property the two competitor metrics lack. The separation holds in 10 of the 12 cells: UEC is negative under shortcut removal in all twelve, and positive under covariate shift in ten, the exceptions being the 400-epoch cells for gradient$\times$input (-0.006) and integrated gradients (-0.019), where the update has overrun its own null. One caveat on fidelity to Eq. (2): the differentiation run subtracts ρ_{null} directly rather than $\max(\nu, \rho_{\text{null}})$ (`baselines.py:65`). The cell quoted uses integrated gradients, which is deterministic and has $\nu = 0$ exactly, so the two agree there.

F.6 The re-audit's resampling

Hemmat & Fatemi (2025) audit 45 settings: five classical learner families, three UCI datasets, three A/B configuration pairs each. Breast cancer is 569×30 over 2 classes, wine 178×13 over 3, digits 1797×64 over 10, each standardised on the full sample (`run_reaudit.py:130--131`). We reimplement their attribution as occlusion in standardised space against a class-anchored margin, averaged over the mean and median baselines (`run_reaudit.py:75--95`), and measure movement with the same $d(\phi(\cdot), \phi(\cdot))$ as above (`run_reaudit.py:109`).

The null is a **bootstrap with replacement**, not a holdout. For each of 10 seeds, `rng.choice(len(X), len(X), replace=True)` draws n indices from the n training rows (`run_reaudit.py:139`); configuration A is refitted on that resample at the same hyperparameters and the same learner seed (`run_reaudit.py:144`), while the treatment refits configuration B on the full sample (`run_reaudit.py:143`). Treatment and null are thus matched in n, in learner and in seed, and differ only in which rows they see. The probe is 200 points drawn without replacement, 178 on wine. That is 10 resamples per setting and 450 null refits in total;

because configuration A coincides across the three pairs for logistic regression, random forests, gradient boosting and k-NN, the 45 settings rest on 18 distinct null specifications. A setting clears the floor when a Wilcoxon signed-rank test over the 10 seeds returns $p < 0.05$ and the mean treatment movement exceeds the mean null movement (`run_reaudit.py:173`).

Learner family	settings	clear the floor	median ratio
gradient boosting	9	4	0.616
k-nearest neighbours	9	3	0.518
logistic regression	9	3	0.860
random forest	9	1	0.889
support vector machine	9	4	0.922
all	45	15	0.861

29 of 45 settings (64%) move less than the floor and 15 (33%) clear it. The remaining one, logistic regression under `l2→l2-weak` on breast cancer, moves 1.09 times the floor but not significantly ($p = 0.19$), so it is in neither class.

F.7 Which way the bootstrap cuts

A bootstrap resample shares rows with the sample that produced the reference fit; two fresh draws share none. The null movement should therefore be smaller than it would be against fresh data, and we measure the gap instead of asserting its direction. Split each dataset into two disjoint halves H_1, H_2 of size $m = \lfloor n/2 \rfloor$. The fresh-draw floor is the movement between fits on H_1 and on H_2, two independent samples of size m; the bootstrap floor is the movement between fits on H_1 and on a bootstrap resample of H_1, also of size m. Sample size, configuration, seed, probe and attribution are identical across the two arms, so only the sharing differs. Over the 18 distinct learner $\times$ dataset specifications at 5 seeds, the bootstrap floor is the smaller one: median ratio 0.747 $[0.682, 0.806]$, smaller in 17 of 18 specifications and in 85 of 90 paired runs, Wilcoxon $p = 1.5 \times 10^{-5}$. The one exception is gradient boosting on wine, at 1.006.

The size of the gap is the size theory predicts. A bootstrap sample reuses a measured 0.638 of its parent's distinct rows, against $1 - e^{-1} = 0.632$, so a bootstrap refit differs from its reference with roughly half the variance of two independent fits, and movement, being a distance, should scale as $1/\sqrt{2} = 0.707$. The observed 0.747 sits just above that. The ratio does not trend with m across the three datasets, at 0.710, 0.768 and 0.694 for $m = 284$, 898 and 89, so it transfers from the half-sample design to the full-sample floor the re-audit uses.

The caveat therefore cuts towards the paper's conclusion rather than against it. The bootstrap sets the *lower* bar, so it is harder for a setting to fall below it, and the reported 64% is a floor on what a fresh-draw null would give. Scaling every null by $1/0.747$ moves the count from 29 to 30 of 45 (67%) and the median ratio from 0.861 to 0.643; scaling by the theoretical $\sqrt{2}$ gives 31 of 45 (69%) and 0.608. The direction is the conclusion, and it is not sensitive to which of the two scalings is used.

G Benchmark against explanation-distribution detection

§2 argues that detecting a shift in the explanation *distribution* under a frozen model is a different measurement from ours rather than a competing one. That is a structural claim, and a reviewer was right to ask for a number. We reimplemented the detector of Mougan et al. (2023) – attributions on source-drawn inputs against attributions on target-drawn inputs, a classifier trained to tell them apart, its AUC as the statistic – using the same cross-fitted logistic estimator the ACS arm already uses, so the comparison is not made with an estimator chosen to suit us.

Two calibrations before any number is read. Their detector is a two-sample test and assumes independent samples. Two checkpoints scored on one probe give matched *pairs*, and feeding those to the estimator depresses its AUC to 0.38–0.42 on data whose true value is 0.5. We therefore take one half of the probe from each checkpoint. Second, the estimator is not unbiased at this sample

size: split a single sample against itself, where the truth is exactly 0.5, and it returns 0.506. Every AUC below is read against that floor rather than against 0.5.

Table 6: Their detector and ours, on the same checkpoints. *Their object* is a frozen model with source- against target-drawn inputs. *Above floor* turns the detector on a model update and reports the excess over its own 0.5 calibration. The probe here is 1,000 points rather than the 500 of §6.3, so the ratios are close to but not identical with those there.

Shift	Their detector, their object	Their detector, above floor	Our ratio
Covariate	0.956	+0.060	1.58
Concept	0.495	+0.008	2.24
Shortcut removal	0.518	+0.009	1.41
None (placebo)	0.495	+0.003	1.07

Their detector works, on its object. Covariate shift returns 0.956. It is also at chance for concept shift and shortcut removal, and that is not a defect: with the model frozen, $E_f(x)$ is a fixed function of x, so moving $P(Y \mid X)$ while leaving $P(X)$ alone changes nothing for it to detect. This is the same fact §2 uses to argue that instance-level change is vacuous in their setting, read from the other side. Their method is built for the covariate case and succeeds there.

Neither instrument subsumes the other. Turned on a model update with the data held fixed, their detector stays within 0.06 of its own floor, because the prediction-preserving updates we study move attributions per instance without moving their distribution far enough for a classifier to separate. Ours reads 1.07 to 2.24 on those same checkpoints. Symmetrically, ours is identically zero in their setting by the argument of §2. The honest summary is that the two answer different questions: theirs asks whether the data moved, ours asks whether the explanation moved further than the update alone would move it. A practitioner who wants both needs both, and reporting one as though it covered the other is the error this appendix exists to rule out, in our direction as much as theirs.

H THE CIFAR ARM, BY SHIFT TYPE

The vision arm carries a specific burden. ω needs a Bayes-optimal predictor in closed form, which Waterbirds, Camelyon17 and ImageNet-C do not supply, so the warranted reference cannot be tested on them at any compute budget. What can be tested at 3072 dimensions is the matched null, against a spurious feature whose ω we know by construction. Two questions then matter: does the shortcut result depend on the shortcut being a token, and does the measure ever return nothing?

The original arm averaged Gaussian noise with contrast and planted a 6×6 corner patch tinted by class parity. A corner patch is a cheap feature for a convolutional network to isolate, and it is not the geometry a natural spurious correlation has: in Waterbirds the confound is the background, spatially dominant and entangled with the object. We therefore added a background shortcut that tints everything outside a central 16×16 box, leaving the region that stands in for the object untouched, and split the corruptions so no single operator can hide inside an average. Three seeds, 48 rows per cell, identical operator throughout.

The shortcut result is not a patch artefact. Moving the spurious feature from a corner token to the background changes the ratio from 1.34 to 1.33 and the exceedance from 0.27 to 0.28. The background arm is also the steadier measurement, at ± 0.06 across seeds against ± 0.12, and it preserves more probe points, so it is better powered rather than merely agreeing. Whatever the network is doing when the shortcut is withdrawn, it does not depend on the shortcut having been cheap to localise.

Brightness returns nothing, and that is the point. A pure brightness shift moves attributions no further than refitting on fresh source data: 1.03, with 6.4% of probe points above a floor defined to admit 5%. Per seed the exceedance is 0.079, 0.070 and 0.044. The measure is therefore not a device that finds unwarranted change wherever it looks. A global brightness offset is affine, and a first layer can absorb it in its biases without reorganising which features carry the prediction, so

Table 7: Each shift type against its own matched null. *Exceedance* is the share of probe points above the null's 95th percentile, so a family indistinguishable from its null sits at 0.05 by construction. The first row is the arm reported in §6.3 and is unchanged; adding the others perturbs neither it nor the corner-patch row.

Shift	Δ	ρ_{null}	Ratio	Exceedance	Reading
Noise + contrast (published)	0.178	0.143	1.37	0.41	present
Contrast alone	0.189	0.143	1.48	0.49	present
Pixelate alone	0.186	0.145	1.42	0.43	present
Gaussian noise alone	0.171	0.143	1.28	0.31	present
Brightness alone	0.145	0.142	1.03	0.06	*at the null*
Shortcut, corner patch	0.152	0.124	1.34	0.27	present
Shortcut, background	0.173	0.143	1.33	0.28	present

there is nothing for the measure to find and it finds nothing. The three corruptions that do reorganise the input – contrast, pixelation, noise – all exceed their nulls.

This is a refinement of §6.3 and not a retraction of it. The arm that section reports is the noise-and-contrast pair, whose numbers are unchanged, and unwarranted change is present there. What the split adds is that the claim is a statement about particular shifts rather than about corruption in general.

I THE EXPONENTIAL-MOVING-AVERAGE ARM

An EMA of the weights is standard practice for stabilising a model across an update, so if it also supplied a floor for explanation movement the matched null would be redundant. We ran it rather than reason about it. `run_ema_control.py` keeps a shadow average of the parameters during the update and returns that point instead of the final iterate. The arm draws from the same stream and runs at the same seed as the arm it smooths, so the optimiser sees an identical minibatch sequence; only the point returned from the trajectory differs. Four families, 5 seeds, two explainers, three decays and both initialisations, at the synthetic headline operator: 240 rows.

Two readings of the suggestion are separable and we report both. *EMA as the reference* takes $\Delta/\Delta_{\text{ema}}$, measuring movement from the source to the smoothed update. *EMA on both arms* takes $\Delta_{\text{ema}}/\rho_{\text{ema}-\text{null}}$, applying the same smoother to treatment and to the matched null and leaving the null as the reference.

Which EMA is on trial. The choice of initialisation is not cosmetic here, so we run both. An average seeded at the starting weights retains decay^T of them after T steps: over the 32 steps of this update that is 3.4% at decay 0.9 but 96.8% at 0.999, so a high-decay arm largely reports the source model rather than the update. This is not a strawman we constructed. It is what `torch.optim.swa_utils` implements, whose documentation defines $W_0^{\text{EMA}} = W_0^{\text{model}}$ and which applies no correction. The alternative, Adam's convention, seeds the average at zero and divides by $1-\text{decay}^T$, giving a weighted mean of the iterates actually visited. We report the corrected arm as the primary result, because the claim should rest on the method at its best.

The corrected arm carries the finding. EMA read as the reference returns 1.25 to 1.72 on data where the answer is 1 by construction, and it does so stably: the number no longer runs away with the decay, so the failure is a property of the method and not of one implementation. The uncorrected arm fails much harder, and the last column says why. Its ratio tracks the warm-up residual almost exactly, because a denominator that has barely left the source model can be made as small as one likes by raising the decay. A reader who reaches for the PyTorch default gets that behaviour, which is worth knowing, but it is not the ground the argument stands on.

EMA does reduce variance, which is the property it is chosen for and which we do not dispute. Corrected, the across-seed standard deviation of Δ falls to 0.832, 0.663 and 0.643 of its unsmoothed value at the three decays. The uncorrected figures, 0.805 to 0.021, overstate this considerably: an

Table 8: Placebo, where source and target are the same environment, so a reference that is working returns 1. The matched null does. The seed floor does not, which is §6.1. Neither does EMA read as the reference: corrected it fails modestly and stably, uncorrected it fails without bound as the warm-up residual grows.

Variant	Decay	Matched null	Seed floor	EMA as reference	EMA on both arms	Warm-up
Corrected	0.9	1.08	0.19	1.25	1.09	—
	0.99	1.08	0.19	1.65	1.10	—
	0.999	1.08	0.19	1.72	1.10	—
Uncorrected	0.9	1.08	0.19	1.30	1.09	3.4%
	0.99	1.08	0.19	5.97	1.10	72.5%
	0.999	1.08	0.19	54.01	1.11	96.8%

average that has not moved is trivially stable, and most of that apparent reduction is the warm-up residual rather than smoothing.

Used on both arms the smoother is harmless and the matched null continues to carry the reference: 1.09 to 1.11 on the placebo across every decay and both variants, against the null's own 1.08, and under shift the ordering is unchanged (covariate 1.67 against 1.57, concept 2.06 against 2.21, shortcut 1.40 against 1.41, corrected at decay 0.99). The two are complementary. Only one of them is a control.

J FIGURE ACCESSIBILITY

Figures use the Okabe–Ito palette (Okabe & Ito, 2008) and encode every distinction redundantly, in hatch for bars and in marker shape for lines, so that they survive monochrome printing and the common colour-vision deficiencies. We simulated deuteranopia, protanopia and tritanopia by cone collapse and checked pairwise separation for the colours that share each panel; where separation on colour alone falls below threshold, the redundant encoding carries the distinction.