EVIDENCE INHERITANCE IN MODEL UPDATES: PERFORMANCE CONTINUITY IS NOT ENOUGH

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployment teams commonly release updated models after data refreshes, continual training, or maintenance, while continuing to cite behavioral evidence from an earlier version. We study *evidence inheritance*: whether diagnostic evidence established for a reference model M_0 remains empirically representative of an updated model M_t. Identification uses a three-level control chain. Level 1 is a magnitude sanity check: positive-scale reparameterizations match the weight-space or representation-space displacement of a claim-breaking update while the prediction function stays fixed, so near-zero diagnostic transfer loss (DTL) is the arithmetic consequence of that symmetry and only closes the false positive that any movement of weights or features would invalidate the diagnostic. Level 2 is the identification. Reliance migrates onto a third channel the inherited diagnostic never probes: the function changes and weight-space displacement exceeds the claim-breaking update, yet the inherited claim remains intact. Level 3 transfers that sanity check, and the claim-breaking phenomenon, to stacked MNIST–CIFAR inputs on a convolutional network. Prediction churn and claim-aligned functional drift (FD) reverse rank between a saturated synthetic task and stacked image inputs. Thus performance continuity is insufficient for automatically inheriting historical diagnostic evidence: continuity must be assessed relative to the historical claim being transferred.

1 INTRODUCTION

Modern deployed models are updated through data refreshes, continual training, compression, and routine maintenance. Yet model cards, subgroup analyses, stress tests, and behavioral diagnostics are normally established for a particular version (Sculley et al., 2015; Mitchell et al., 2019; Raji et al., 2020). Suppose a team has established behavioral evidence for M_0, releases M_t, and observes essentially unchanged accuracy. Which parts of that evidence, if any, still describe M_t?

This differs from whether the new model is worse. Even when $P(M_t) \approx P(M_0)$, similarly accurate models can follow different rules (Fisher et al., 2019; Marx et al., 2020; D'Amour et al., 2022), and average accuracy does not fully describe cross-version compatibility (Milani Fard et al., 2016; Träuble et al., 2021; Yan et al., 2021). We call this longitudinal question **evidence inheritance**: whether a version-indexed diagnostic claim established for M_0 remains empirically representative of M_t. Prediction compatibility compares outputs; monitoring observes inputs or aggregate error; regression testing checks predefined tests. Evidence inheritance asks whether a historical *claim* still applies. Functional drift (FD) measures claim-aligned behavior change; diagnostic transfer loss (DTL) is the independent diagnostic endpoint.

Our contributions.

1. **Problem formulation.** We make the cross-version empirical representativeness of a historical diagnostic claim explicit, separating it from prediction compatibility, shift monitoring, and regression testing.
2. **Three-level identification.** Level 1 is a magnitude sanity check: positive-scale rewrites matched in weight or representation displacement keep the prediction function fixed, so near-zero DTL is analytic. Level 2 is the identification: reliance moves onto a channel the claim does not probe, weight-space displacement exceeds the claim-breaking update, and the inherited claim

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

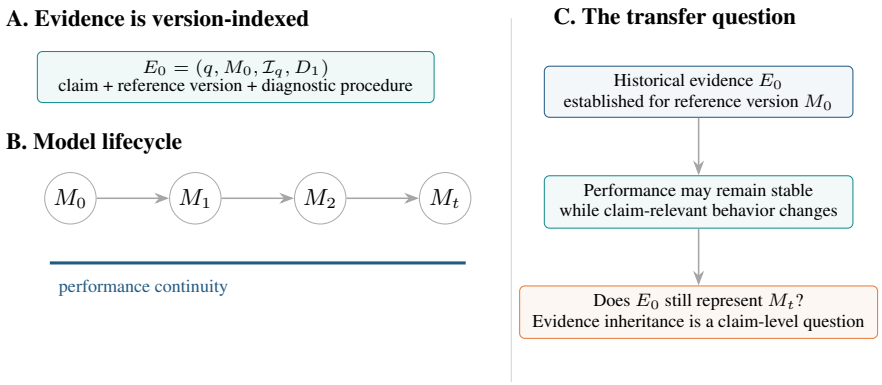


Figure 1: Evidence inheritance under model updates. Historical evidence binds a reference model, a diagnostic procedure, interventions, and a claim. A later version may preserve ordinary performance while changing claim-relevant behavior and its independent diagnostic consequence. The question is whether E_0 still represents M_t.

remains intact. Level 3 transfers the sanity check and the claim-breaking phenomenon to stacked MNIST–CIFAR inputs on a convolutional network.

3. **Scope and measurement boundary.** A preregistered second construction removes the first construction's channel asymmetry. Prediction churn and FD reverse rank between a saturated synthetic task and stacked image inputs. On the orthogonal construction, only claim-aligned FD_q separates claim-breaking from claim-benign function change.

2 Evidence Inheritance under Model Updates

2.1 Evidence object and diagnostic claim

Let q be a historical claim about model behavior, $\mathcal{I}_q$ its prespecified intervention family, and D_1 a diagnostic sample. We instantiate historical evidence as

$$E_0 = (q, M_0, \mathcal{I}_q, D_1, \widehat{R}_{D_1}(M_0), \widehat{G}_0). \tag{1}$$

The same binding can cover calibration, subgroup risk, or human-evaluation claims; we instantiate one diagnostic-risk object. $\widehat{G}_0 : \widehat{R}_{D_1}(M_0) \le \epsilon$ with $\epsilon = 0.05$ is a finite-sample eligibility statement. Evidence refresh means rerunning the procedure for M_t; evidence inheritance means attaching the old evidence to M_t without that new measurement (Mitchell et al., 2019; Collins et al., 2024).

2.2 Three distinct quantities

Performance continuity is only an ordinary-task constraint,

$$\text{PerfCont}_t : \quad |P(M_t) - P(M_0)| \le \delta_P. \tag{2}$$

Claim-relevant functional migration is a measurable change in the model response to $\mathcal{I}_q$. The evidence consequence is evaluated separately, on a diagnostic sample $D_2 \sim \mathcal{D}_{\mathcal{I}}$:

$$\widehat{\text{DTL}}_t = \widehat{R}_{D_2}(M_t) - \widehat{R}_{D_2}(M_0). \tag{3}$$

Positive DTL means higher risk under the old diagnostic intervention family. Inheritance is the subsequent version-management judgment, and it also depends on the reference margin.

2.3 A claim-aligned behavioral measurement

To operationalize whether behavior relevant to the historical claim has changed, we construct a minimal intervention-response measurement, FD. For $F_q(M) = [I_{q,A}(M), I_{q,B}(M)]$,

$$I_k(M) = \mathbb{E}_{x \sim D_{\text{FD}}} \left[|p_M(x) - p_M(\iota_k(x))| \right], \qquad \widehat{\text{FD}}_t^{(q)} = \|\widehat{F}_q(M_t) - \widehat{F}_q(M_0)\|_2. \tag{4}$$

Table 1: Four-arm design. The weight-matched and representation-matched arms are the magnitude sanity check: the prediction function is held fixed, so near-zero DTL is the arithmetic of that symmetry. The orthogonal control is Section 4.2. FD, DTL, and independent outcome data are not used to select matches.

Arm	Construction goal	Rival explanation controlled	Expected consequence
MATERIAL migration	Move reliance from A to B	Whether the phenomenon exists	Positive DTL
Matched-WD	Match MATERIAL weight distance by the positive-scale symmetry	Displacement without function change	Near-zero DTL (analytic; magnitude sanity check)
Matched-RD	Match MATERIAL representation distance by the same symmetry	Displacement without function change	Near-zero DTL (analytic; magnitude sanity check)
Natural benign	Ordinary warm start	Update or stage alone	Benign DTL (descriptive)

The signature encodes responses to interventions specified by the old claim, so the same pair of models can have small drift for one claim and large drift for another. Interventions are frozen before outcomes are observed. For a risk-upper-bound claim, a positive DTL that exceeds the reference margin fails the frozen transfer rule. In deployment, claims, protected slices, or counterfactual operations are extracted from the original validation materials and frozen as $\mathcal{I}_q$ before the update is inspected.

3 CONTROLLED COMPARISON STRATEGY

3.1 A THREE-LEVEL CONTROL CHAIN

A correlation between a changing signature and diagnostic risk does not establish evidence materiality: both may merely track training stage, or the claim-breaking update may simply move farther. We therefore use a three-level control chain. Level 1 is a *magnitude sanity check*. The weight-matched and representation-matched arms (Matched-WD and Matched-RD) apply the same positive-scale ReLU symmetry, one selected to match the weight distance of the claim-breaking (MATERIAL) update and one selected to match its representation distance. The prediction function is unchanged in exact arithmetic (prediction difference $\leq 10^{-5}$), so near-zero DTL and prediction churn are the arithmetic consequence of that symmetry. A diagnostic that failed whenever weights or representations moved would assign large DTL to these arms. Natural is an ordinary warm start. Level 2 (Section 4.2) is the identification: two function-changing updates, one claim-breaking and one claim-benign, isolate channel migration from function change as such. Level 3 transfers the sanity check and the claim-breaking phenomenon to stacked image inputs (Section 4.3).

The four-arm MAE table records this sanity-check pool: distance predictors are large on the symmetries, where DTL is zero because the outputs are unchanged, while FD stays near zero. Identification of two function-changing updates is Table 4.

3.2 DATA, MODEL, AND ISOLATION

Labels are $Y \sim \text{Bernoulli}(1/2)$, $S_Y = 2Y - 1$, and two channels are generated as

$$X_A = S_Y(1 - 2Z_A) + \varepsilon_A, \qquad X_B = S_Y(1 - 2Z_B) + \varepsilon_B, \tag{5}$$

$$\varepsilon_A, \varepsilon_B \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.05^2). \tag{6}$$

The MATERIAL schedule changes (r_A, r_B) from $(0.98, 0.80)$ toward $(0.90, 0.99)$; the ordinary performance environment keeps both channels highly predictive. Thus observational performance can remain similar while the frozen diagnostic intervention, which degrades B, separates solutions. The construction is a controlled stress test (Geirhos et al., 2020; D'Amour et al., 2022).

Table 2: Frozen measurements in the four-arm synthetic construction. Each quantity is computed without using diagnostic-outcome labels.

Measurement	Definition	Evaluation object
ΔP	$\lvert\widehat{R}_{\mathrm{perf}}(M_t) - \widehat{R}_{\mathrm{perf}}(M_0)\rvert$	D_{perf}
WD	$\lVert\theta_t - \theta_0\rVert_2 / \lVert\theta_0\rVert_2$	parameters
RD	$1 - \mathrm{CKA}(H_t, H_0)$	frozen probe
FD	$\lVert\widehat{F}(M_t) - \widehat{F}(M_0)\rVert_2$	D_{FD}

We train a $2 \to 32 \to 1$ ReLU–sigmoid MLP from a reference model through ten warm-start stages. D_1, D_2, and D_{FD} are independent frozen pools: D_{FD} estimates the claim-aligned signature, D_2 evaluates diagnostic consequence, and matching sees neither FD nor D_2 outcomes (Appendix A.2).

3.3 Metrics and inference

We compare FD with performance change ΔP, normalized weight distance (WD), and representation distance $\mathrm{RD} = 1 - \mathrm{CKA}(H_t, H_0)$ (Kornblith et al., 2019). Each predictor is mapped to checkpoint DTL with the same linear calibration protocol. On the four-arm pool this protocol is the sanity-check comparison: 5-fold seed-held-out cross-validation, seed-level bootstrap intervals, and paired sign-flip tests. Eligibility and matching gates are specified before that comparison (Table 8). The identification comparison is Section 4.2.

3.4 Matching and reporting hierarchy

For every Material checkpoint, a Matched-WD and Matched-RD control is selected from a prespecified candidate set using only its target distance,

$$c^* = \arg\min_{c \in \mathcal{C}_{s,t}} |d_b(M_0, c) - d_b(M_0, M_t^{\mathrm{material}})|. \tag{7}$$

Candidate selection never uses FD, DTL, or the D_2 outcome sample. Both matched arms use the positive-scale ReLU reparameterization in Appendix A.2; they differ only in the distance they are asked to match. Natural uses warm starts with no directed migration. Enumeration, scales, and thresholds are in Appendix A.2–A.4. Seeds are the inference units. Checkpoint-level MAE on the four-arm pool is the sanity-check comparison; pairwise differences are aggregated within seeds.

4 Three-Level Identification

We first record the phenomenon, then identify it with the three-level chain. Of 200 seeds, 187 established the reference claim and entered the primary analysis. Eligible Material trajectories stayed below the two-percentage-point performance-continuity threshold (median within-seed accuracy range 0.30 points). All 187 show $I_A \downarrow, I_B \uparrow$. Median final DTL is 20.2 percentage points versus a 2.0 point median reference margin.

4.1 Level 1: Magnitude sanity check

The weight-matched and representation-matched arms attain 100% and 95.2% frozen coverage. Both keep the prediction function fixed (prediction difference $\leq 10^{-5}$), so final DTL and prediction churn sit at numerical zero, against 20.2 percentage points of DTL for Material (Figure 3). A rewrite at the matched weight or representation displacement leaves the diagnostic intact. Because those arms preserve outputs in exact arithmetic, that zero is the symmetry, and it is the whole content of Level 1.

The four-arm pool used for the preregistered comparison mixes Material with these symmetries. On the matched arms, WD and RD follow the claim-breaking trajectory while DTL, FD, and prediction churn are numerical zeros. A linear map from WD or RD to DTL is therefore charged for large distances paired with zero loss, and FD's lower held-out MAE is the expected score of that pool: 0.0157 versus 0.0170 for ΔP, 0.0328 for WD, and 0.0232 for RD ($p = .00186$, $p < 10^{-5}$, and

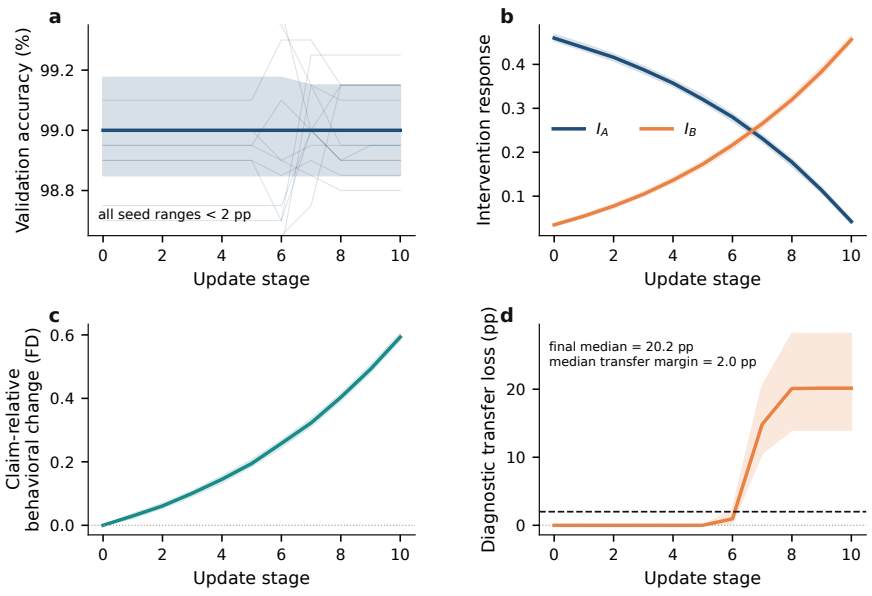


Figure 2: Performance continuity can conceal claim-relevant functional migration ($n = 187$ eligible seeds). Ordinary accuracy stays stable, reliance migrates from A to B, FD increases, and independent diagnostic DTL rises after migration. Lines are seed medians and bands are IQRs.

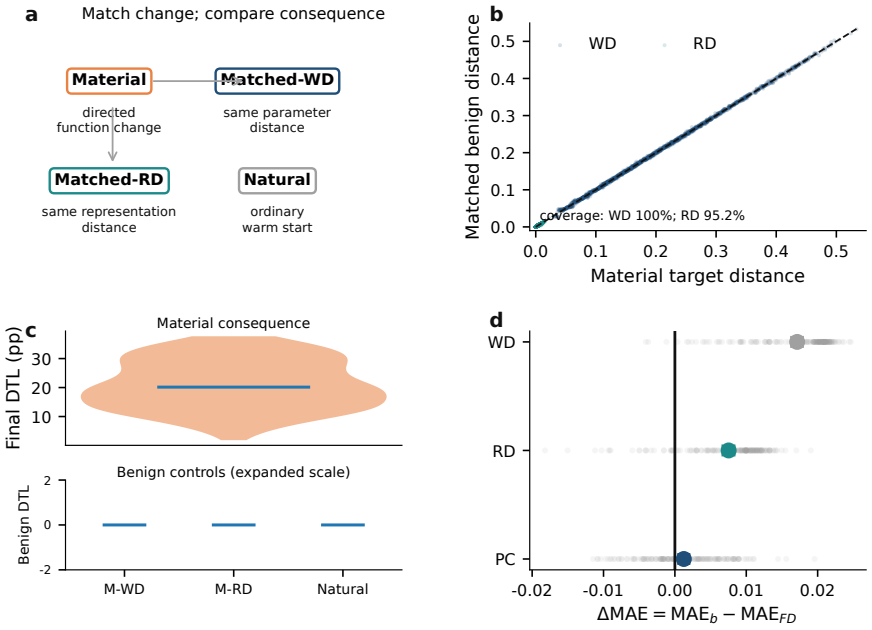


Figure 3: Level 1 magnitude sanity check ($n = 187$): positive-scale matches reproduce target weight- and representation-distance scales. DTL stays at zero because the prediction function is unchanged. The held-out panel records error on this sanity-check pool.

$p < 10^{-5}$). The preregistered intersection–union criterion records this sanity check. The comparison of two function-changing updates is Table 4.

Refits on the MATERIAL arm alone keep the same ordering with a smaller gap (Appendix A.4). Training only on the benign arms and testing on MATERIAL yields MAE 0.0732 for every candidate. Natural is an ordinary warm start, and a positive FD can accompany zero DTL.

Table 3: Sanity-check pool ($n = 187$): held-out error with exact reparameterizations included. On those arms DTL is zero because the prediction function is unchanged, so the gaps against WD and RD are the arithmetic of that pool. Identification: Table 4.

Comparison	ΔMAE	95% CI	p
FD vs. ΔP	0.00125	[0.00048,0.00203]	.00186
FD vs. WD	0.01713	[0.01627,0.01796]	$< 10^{-5}$
FD vs. RD	0.00753	[0.00666,0.00837]	$< 10^{-5}$

4.2 Level 2: Orthogonal control

Level 1 fixes the prediction function, so its near-zero DTL is the symmetry. Level 2 changes the function. We add a third channel C that the inherited diagnostic never permutes, and migrate reliance from A to C rather than to B. Only B-degradation is diagnostic. Claim-aligned $\mathrm{FD}_q = |I_B(M_t) - I_B(M_0)|$. When the function changes, the diagnostic remains valid if the change avoids the diagnostic channel, isolating channel migration as the cause of evidence failure.

Of 200 independent seeds, 191 establish evidence and remain performance-continuous. Every orthogonal trajectory changes the function (median final max-probability difference 0.925) while remaining claim-benign (median final DTL -1.9 percentage points; maximum $+0.45$). Material updates raise median final DTL to 23.1 percentage points. Orthogonal WD is larger than Material WD (0.566 versus 0.452), so a larger parameter displacement coincides with a benign claim. Median final representation distance stays small on both function-changing arms (0.007 on Material, 0.011 on the orthogonal arm), so Level 2 identifies the claim for weight displacement; large representation displacement with an unchanged function remains the Level 1 symmetry. Ordinary-validation prediction churn is identical at the median (0.0195) on both function-changing arms; the two-channel signature FD rises on both because I_A drops. Only FD_q stays small on the orthogonal arm (0.021 versus 0.424). Accordingly, FD_q attains lower held-out MAE than churn (0.0235 versus 0.0298; $\Delta\mathrm{MAE} = 0.00633$, 95% CI $[0.00548, 0.00714]$, $p < 10^{-5}$). Unindexed two-channel FD is worse than churn (MAE 0.0377). Specificity therefore requires indexing the measurement to $\mathcal{I}_q$, not merely detecting that the function moved.

Table 4: Orthogonal moving control on a three-channel construction ($n = 191$). Medians at the final stage. Orthogonal weight distance exceeds Material. Median final representation distance is 0.007 on Material and 0.011 on the orthogonal arm, so this table identifies weight displacement. Prediction churn is measured on ordinary V_S.

Arm	DTL	WD	FD_q	FD	Churn
Material $A \to B$	0.231	0.452	0.424	0.596	0.0195
Orthogonal $A \to C$	-0.019	0.566	0.021	0.438	0.0195
Natural	0	0.351	0.006	0.010	0

4.3 Level 3: Transfer to stacked image inputs

Level 3 asks whether the sanity check and the claim-breaking phenomenon still appear on a convolutional network. The construction follows the Dominoes design of vertically stacking two image patches (Izmailov et al., 2022). Each input is a 64×32 RGB mosaic: channel A is MNIST digits $0/1$ (LeCun et al., 1998) and channel B is CIFAR-10 automobile/frog (Krizhevsky, 2009); a diagnostic intervention permutes only the corresponding 32×32 block. The model is a three-layer convolutional network without normalization, so the weight-matched arm remains an exact ReLU-homogeneous channel rescaling. Because linear CKA is invariant to orthogonal transforms, the representation-matched arm uses an invertible probe mix $h' = hA$ with head compensation $W' = WA^{-T}$. Thresholds $\epsilon = 0.07$ and $\tau = 0.04$ were set on a disjoint calibration set. Of 200 seeds, 174 are primary-eligible. Median final DTL is 0 on the weight- and representation-matched arms, the same arithmetic consequence of an exact reparameterization as in Level 1, and 0.149 on Material.

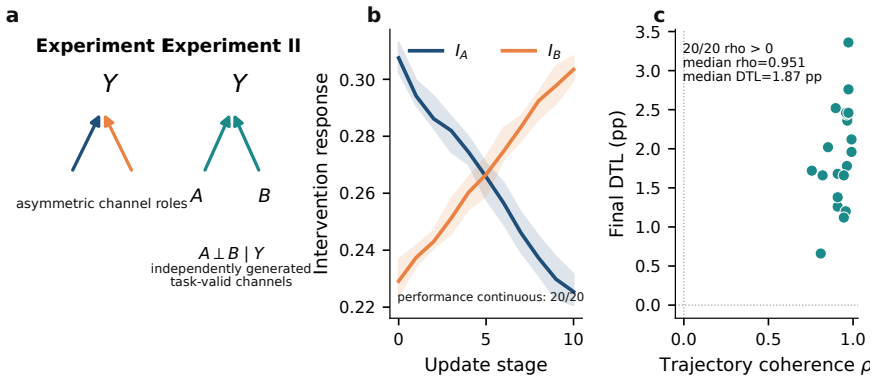


Figure 4: Preregistered scope check with two independently generated task-valid channels. The pattern persists without a designated shortcut.

Table 5: Sanity-check pool on stacked image inputs ($n = 174$ eligible seeds). The matched arms are exact reparameterizations, so the gaps against WD and RD repeat the Level 1 arithmetic. $\Delta\text{MAE} > 0$ favors FD inside that pool.

Comparison	ΔMAE	95% CI	p
FD vs. ΔP	0.00714	[0.00679, 0.00748]	$< 10^{-5}$
FD vs. WD	0.00725	[0.00686, 0.00761]	$< 10^{-5}$
FD vs. RD	0.00766	[0.00727, 0.00804]	$< 10^{-5}$

The weight- and representation-matched arms remain exact reparameterizations, so a median final DTL of 0 is the same analytic outcome as Level 1. The held-out gaps in Table 5 score distance predictors on that symmetry. The image quantity independent of the symmetry is the claim-breaking arm, with median final DTL 0.149. Table 6 uses an independent seed block.

5 Scope Check: Two Task-Valid Channels

The four-arm synthetic construction leaves a specific alternative mechanism: perhaps its consequence is entirely caused by a prespecified migration from a stable channel to an environment-sensitive channel. This preregistered scope check tests only whether that asymmetry is necessary. It uses two independently generated, task-valid predictive channels, neither designated as spurious or shortcut, with $Y \to A$, $Y \to B$, and $A \perp B \mid Y$.

All 20 preregistered seeds show $I_A \downarrow, I_B \uparrow$, median $\rho(\text{FD}, \text{DTL}) = .951$, and median final DTL 1.87 percentage points. Calibration was frozen before evaluation seeds were drawn (Appendix A.4).

6 Measurement Boundary: No Universal Single Metric

Could the previous results be read as evidence that FD is a generally preferred drift measure? We tested that interpretation on two carriers, collecting prediction Churn, probability and logit distances, and claim-aware behavioral drift (CABD) during training. None of the candidates reads DTL labels. The four-arm image transfer in Section 4.3 uses an earlier frozen seed block; the image column below uses a later independent block.

On a saturated synthetic replication ($n = 193$), Churn beats FD because every benign arm has zero prediction flips (Table 6). On stacked image inputs the natural-arm Churn median is 0.009 (max 0.171; $n = 166$), and FD then beats Churn (0.0106 versus 0.0170; $\Delta\text{MAE} = 0.00639$, 95% CI $[0.00590, 0.00688]$, $p < 10^{-5}$). The ranking reverses once benign updates leave the zero-flip regime. Nested comparisons are in Appendix A.5.

Table 6: Two-carrier behavioral baselines (synthetic $n = 193$; image $n = 166$). Same classification-risk DTL; none of the candidates uses DTL labels. Bold marks the lower MAE on that carrier.

Measurement	Synthetic MAE	Image MAE
Churn	**0.00698**	0.01698
FD	0.01617	**0.01059**
CABD	0.01636	0.01493
ΔP	0.01742	0.01777
ProbDist	0.02351	0.01790
RD	0.02441	0.01819
WD	0.03511	0.01783

Table 7: Object hierarchy of neighboring problems: what is transferred, what conditions the comparison, and what judgment follows.

Lineage	Object compared or transferred	Conditioning information	Typical judgment
Prediction compatibility	Per-example predictions or error sets	Inputs and reference outputs	Whether the new version preserves old output behavior
Distribution monitoring	Input, representation, or aggregate-error distributions	Time, environment, or monitoring window	Whether a distributional alarm warrants investigation
Regression / behavioral testing	Predefined tests or specs and pass/fail status	Test set, slice, or specification	Whether the new version still passes established tests
Claim-level evidence inheritance (this paper)	A version-indexed diagnostic claim	Claim, intervention family, reference version, and margin	Whether a specified historical claim remains empirically representative

7 Related Work

Neighboring literatures differ in the object that is compared or transferred (Table 7). Predictive multiplicity, model reliance, and underspecification show that similar accuracy need not determine function (Fisher et al., 2019; Marx et al., 2020; D'Amour et al., 2022). Training randomness is another source of multiplicity (Dodge et al., 2020; Bouthillier et al., 2021; Sellam et al., 2022). Backward-compatible updates, prediction churn, and regression testing transfer outputs or predefined test status (Milani Fard et al., 2016; Yan et al., 2021; Träuble et al., 2021; Bansal et al., 2019); a few prediction flips need not void a claim with sufficient margin, and an old test can still run without its summary conclusion attaching to the new model. Shift detection, explanations, and external validation provide alarms or finer observations (Hendrycks & Gimpel, 2017; Rabanser et al., 2019; Ovadia et al., 2019; Koh et al., 2021; Ribeiro et al., 2016; Kornblith et al., 2019; Collins et al., 2024; Miller et al., 2021), but they become evidence-continuity measurements only when aligned with the old claim and intervention family.

8 Discussion

The results support two asymmetric statements: displacement of an unchanged prediction function leaves the diagnostic intact, and performance continuity leaves open whether a historical claim still represents the update. The three-level chain separates those statements. Level 1 shows that positive-scale reparameterizations matched in weight or representation distance leave DTL at zero because the function is fixed. Level 2 shows that a function-changing update can move farther in weight space than the claim-breaking update and still leave the inherited claim intact when the migration avoids the diagnostic channel; representation distance on those trajectories stays small. Level 3 carries the sanity check and the claim-breaking phenomenon to stacked image inputs. Continuity may therefore vary across claims after one global update. Which measurement is most informative depends on whether benign updates occupy a zero-flip regime, and on whether the measurement is indexed to $\mathcal{I}_q$.

9 Limitations

Both constructions use low-dimensional systems and one point-estimate diagnostic-risk claim. The image setting is a two-block mosaic; the orthogonal control adds a third synthetic channel. Deploy-

ment rules, prevalence in real version histories, and automatic intervention discovery remain open. Claims, interventions, and margins must be frozen before outcomes are observed.

10 Conclusion

We reframe continuity in model updates from "how much did the model change?" to "which evidence established for the reference version still represents the update?" Stable performance is insufficient for automatically inheriting historical diagnostic evidence. A magnitude sanity check shows that function-preserving displacement leaves DTL at zero because the prediction function is unchanged. An orthogonal control shows that channel migration, with a larger weight-space displacement and a changed function, breaks the inherited claim when the migration lands on the diagnostic channel and leaves the claim intact when it does not. The sanity check and the claim-breaking phenomenon transfer to stacked image inputs. Prediction churn and FD reverse rank once benign updates leave a zero-flip regime. Evidence inheritance is therefore a claim-relative longitudinal question.

Reproducibility Statement

Data-generating process, model, sample isolation, matching, and estimands are specified above; appendices record hyperparameters, eligibility gates, nested tests, and the three-channel protocol. The four-arm synthetic construction, the image transfer, both baseline blocks, and the orthogonal control use disjoint seeds. The supplement will include code, seed-level outputs, and figure data.

AI Use Statement

Language-model tools were used for structure discussion, language editing, and terminology checks. Authors are responsible for the experimental design, execution, analysis, and scientific conclusions.

References

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2429–2437, 2019. doi: 10.1609/aaai.v33i01.33012429.

Xavier Bouthillier, Pierre Delaunay, Moustapha Bronzi, Alexey Trofimov, Boris Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Gary S. Collins, Paula Dhiman, Jie Ma, Michael M. Schlussel, Lucinda Archer, Ben Van Calster, Frank E. Jr. Harrell, Glen P. Martin, Karel G. M. Moons, Maarten van Smeden, Matthew Sperrin, Garrett S. Bullock, and Richard D. Riley. Evaluation of clinical prediction models (part 1): From development to external validation. *BMJ*, 384:e074819, 2024. doi: 10.1136/bmj-2023-074819.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *Transactions of the Association for Computational Linguistics*, 8:729–744, 2020.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(180):1–81, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–672, 2020.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38516–38532, 2022.

Pang Wei Koh, Shiori Sagawa, Kai Xiao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

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

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

Charles T Marx, Flávio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6765–6774. PMLR, 2020.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

John P. Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, et al. MultiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*, 2022.

Frederik Träuble, Julius von Kügelgen, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Peter V Gehler. Backward-compatible prediction updates: A probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Shen Yan, Dong Zhang, Yang Yang, and Mingkui Tan. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1424–1433, 2021.

A Supplementary Protocol and Results

A.1 Execution and data integrity

The four-arm synthetic construction uses 200 seeds. Training, evidence construction, consequence evaluation, and functional-response measurement use mutually independent samples. The synthetic baseline, four-arm image transfer, image baseline, and orthogonal-control experiments use disjoint seed blocks. Detailed tables accompany the anonymous supplement.

A.2 Implementation and candidate enumeration

Each seed's reference model is a $2 \to 32 \to 1$ ReLU–sigmoid MLP trained with Adam and binary cross-entropy for 80 epochs (learning rate 10^{-2}, batch size 256), then warm-started at learning rate 10^{-3} for ten update stages of 20 epochs each. The pools D_1, D_2, and D_{FD} each have $n = 2000$; the representation probe has $n = 512$. Training, evidence construction on D_1, consequence evaluation on D_2, functional-response measurement on D_{FD}, and the representation probe use mutually independent random samples. Five-fold splits are seed-level: all arms and checkpoints of a seed remain in the same fold, and standardization is estimated from the training fold only.

Both Matched-WD and Matched-RD use the positive-scale ReLU reparameterization

$$(W_{1j}, b_{1j}, W_{2j}) \mapsto (c_j W_{1j}, c_j b_{1j}, W_{2j}/c_j), \qquad c_j > 0,$$

which changes parameter and hidden-representation distances while preserving network outputs in exact arithmetic. The two arms differ only in the target distance. For every target, the matching system enumerates 8 frozen directions and 81 scales, hence 648 candidate configurations. Selection is deterministic and distance-only; ties are broken by smaller scale, then smaller direction index. Candidates, matching, FD, DTL, and the consequence sample remain isolated; benign consequences are inspected only after matches are locked.

A.3 Image-carrier coverage

Of 200 independent image-carrier seeds, 193 establish evidence at M_0 and 174 of those are performance-continuous; WD coverage is $200/200$ and RD coverage is $193/200$. Natural-arm diagnostic change stays within $\tau = 0.04$ for $191/200$ seeds.

A.4 Eligibility, matching, and secondary analyses

Of 200 seeds, 187 established the reference claim; all 187 remained performance-continuous. Matching coverage, function preservation, and consequence separation hold under the prespecified tolerances in Table 8. The primary comparison is not used to decide whether these conditions hold.

Table 8: Prespecified eligibility and matching criteria for the four-arm synthetic construction.

Criterion	Definition
Evidence availability	At least 80% of seeds satisfy $\widehat{R}_{D_1}(M_0) \le .05$
Performance continuity	At least 80% of eligible seeds have within-seed range ≤ 2 percentage points
Matching coverage	WD and RD each require coverage $\ge .80$ on at least 80% of seeds
Matching error	Median normalized error $\le .10$; 90th percentile $\le .25$
Function preservation	Benign checkpoints: prediction difference, FD, and $\lvert\mathrm{DTL}\rvert$ at most 10^{-5}, 10^{-4}, and .01
Consequence separation	Material DTL range $\ge .05$; matched-arm median max $\lvert\mathrm{DTL}\rvert < .02$
Natural-arm check	At least 90% of seeds have max $\lvert\mathrm{DTL}\rvert$ and range both $\le .02$

When the linear models are refit on the Material arm only, FD has MAE 0.0479, below 0.0601, 0.0544, and 0.0531 for ΔP, WD, and RD; the descriptive R^2 is 0.650 and Spearman ρ is 0.879. The smaller separation relative to the pooled four-arm analysis is what remains after the exact reparameterizations are removed. On those arms DTL is zero by the symmetry, so part of the pooled

gap is that arithmetic. Leave-one-arm-out and natural-update trajectories are supplied with the results package. The identification comparison is the orthogonal control.

For the two-channel scope check, calibration was completed and the specification frozen before evaluation seeds were drawn; no parameter was changed after those seeds began. A non-saturated contrast $(\mu_{\text{strong}}, \mu_{\text{weak}}) = (0.50, 0.40)$ and diagnostic noise grid $\sigma_z \in \{1.25, 1.5, 2.0\}$ were chosen before candidate-specific DTL values were revealed.

A.5 CHURN+FD COMBINATION ANALYSES

The synthetic baseline replication uses a 200-seed block disjoint from the four-arm synthetic construction; 193 seeds establish reference evidence and enter the comparison. Churn, ProbDist, LogitDist, and CABD are recorded during training, so per-example behavior is not reconstructed after the fact from the primary four-arm aggregates. All univariate models use the same seed-held-out protocol as the primary analysis. Table 6 records the two-carrier ranking. On the synthetic carrier, pooled MAE is 0.006981 for Churn and 0.016173 for FD; FD remains below ProbDist, RD, and WD, and is close to CABD.

After observing that comparison, a supplementary analysis on the same 193 eligible seeds compares Churn alone with Churn+FD. Pooled MAE is 0.006981 versus 0.007613, so

$$\Delta\text{MAE} = \text{MAE}_{\text{Churn}} - \text{MAE}_{\text{Churn+FD}} = -0.000632.$$

The seed-bootstrap 95% interval is $[-0.000689, -0.000573]$; the plus-one-corrected paired sign-flip p is 9.9999×10^{-6}; and the relative MAE change is -9.06%. The negative sign means that adding FD increases prediction error. This nested test was not preregistered.

The image-carrier baseline uses a later 200-seed block, disjoint from the four-arm synthetic construction, from the synthetic replication, and from the four-arm image transfer in Section 4.3. Of 200 seeds, 188 establish reference evidence and 166 remain performance-continuous. H1 is the univariate comparison of FD against Churn and is supported in the main text. A preregistered secondary nested comparison on the same 166 seeds compares Churn alone with Churn+FD. Pooled MAE is 0.01698 versus 0.00770, so

$$\Delta\text{MAE} = \text{MAE}_{\text{Churn}} - \text{MAE}_{\text{Churn+FD}} = 0.00927.$$

The seed-bootstrap 95% interval is $[0.00858, 0.00998]$; the plus-one-corrected paired sign-flip p is 9.9999×10^{-6}; and the relative MAE reduction is 54.6%. Adding FD therefore reduces error on the image carrier, the opposite of the synthetic nested result. This nested test is not H1. On the material arm, Spearman correlations with DTL are 0.924 for FD, 0.461 for Churn, and -0.237 for ordinary prediction difference.

A.6 ORTHOGONAL $A \to C$ CONTROL

The orthogonal experiment uses a $3 \to 32 \to 1$ MLP and a third SCM channel C generated like A and B. Diagnostic D1/D2 inherit the four-arm synthetic (A, B) family and keep $r_C = 0.99$; interventions permute A or B only. Material follows the four-arm $A \to B$ schedule with $r_C = 0.50$ fixed; orthogonal uses the same r_A drop with r_B fixed and $r_C : 0.50 \to 0.99$; natural is the usual warm start. Ordinary V_S has $r_A = r_B = r_C = 0.99$. Claim-aligned $\text{FD}_q = |I_B(t) - I_B(0)|$. Seeds 16100–16299 are disjoint from every previously reserved block. Of 200 seeds, 191 establish evidence (all of those remain performance-continuous). All 191 orthogonal trajectories satisfy the function-changing and signed-DTL-benign gates. Median final WD is 0.452 on Material and 0.566 on Orthogonal. Median final RD is 0.007 on Material and 0.011 on Orthogonal. Held-out MAE is 0.0235 for FD_q and 0.0298 for Churn.