A PRINCIPLED DECOMPOSITION OF EXPLANATION FAILURE IN ROUTED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature scores in routed models can disagree with the model's response even when every prediction term is exposed. We decompose this failure for frozen-route gradient-times-input scores under a declared replacement into reference-origin mismatch, finite-step expert approximation, and omitted routing response. The exact decomposition connects each term to a correction or control. Centering removes origin mismatch and affine experts remove expert approximation. Routing can still reverse the response without a support switch. Across ten tabular datasets, five seeds, and four routed predictors, centering increases joint donor-replacement probability drops in 199/200 runs. Complete differentiation improves numerical-field sign agreement in 377/400 checkpoint–reference pairs, but finite-step errors remain. This motivates routing-derivative priorities for budgeted endpoint checks. An original-author SENN reproduction on three datasets extends the coefficient–response distinction to native explanations. Complete derivatives improve response agreement, while a logit-aligned training control incurs a predictive cost. The framework identifies which failures each correction addresses and why local agreement does not guarantee finite-response fidelity.

1 INTRODUCTION

Interpretable expert mixtures expose selected experts and their prediction rules (Ismail et al., 2023; Swamy et al., 2025). Recent work also designs sparse expert activations for intrinsic interpretability (Yang et al., 2025). Yet a routed predictor can expose every prediction term while exporting feature scores that point in the wrong direction when a field changes. Knowing how a prediction was assembled does not establish how the complete predictor responds to a replacement.

Removal-based explanations require a declared replacement and response target (Covert et al., 2021), and feature dependence can make replacements difficult to interpret (Hase et al., 2021). We study disagreement between a feature score and the complete predictor's finite logit response to a specified field replacement. Three sources contribute: gradient-times-input scores depend on a coordinate origin, expert gradients approximate finite expert responses, and changing a field can change routing. Their signed effects can cancel, so an aggregate fidelity score cannot identify the appropriate correction.

Our mechanism evidence routing network (MERN) framework measures these sources on the same predictor and replacement. **(1) Exact error accounting.** We decompose score–response disagreement and establish controlled limits under compensated translations and affine experts (Section 2). **(2) Mechanism-specific corrections.** We distinguish centering, complete differentiation, and finite endpoint evaluation, and motivate a routing-derivative priority for a limited query budget. **(3) Controlled evidence and transfer.** Ten datasets and four routed predictors test these distinctions. Dense affine mixtures isolate routing failures without expert nonlinearity or support switches. An original-author SENN reproduction tests the coefficient–response distinction in an external model.

The algebraic ingredients and corrections build on established ideas. Our contribution organizes them into computable failure terms and testable interventions. The account is specific to the reference and frozen-route convention. Prediction accuracy and correctness ranking are separate evaluation targets.

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

2 A Decomposition of Feature-Explanation Failure

Let $\ell(x) = \sum_e z_e(x) r_e(u)$, with $u = \psi(x)$, $z_e \geq 0$, and $\sum_e z_e = 1$. The expert ledger $z_e r_e$ reconstructs the logit exactly. Each field has one expert scalar $u_j = \psi_j(x_j)$. Routing may depend on all fields. Replacing field j gives x', $u' = \psi(x')$, and $z' = z(x')$, changing only u_j in the expert stream. The target is the finite logit response $D_j = \ell(x) - \ell(x')$ to this declared replacement. It is reference-dependent and need not coincide with other notions of feature importance. Define

$$g_j = \sum_e z_e(x) \partial_j r_e(u), \qquad A_j = u_j g_j, \qquad C_j = (u_j - u'_j) g_j, \quad (1)$$

$$F_j = \sum_e z_e(x)[r_e(u) - r_e(u')], \quad R_j = \sum_e [z_e(x) - z_e(x')] r_e(u'). \quad (2)$$

A is the frozen-route gradient-times-input score and C its reference-centered counterpart. F evaluates the finite expert response with the original route fixed, while $D = F + R$ also updates routing. For categories, the derivative is with respect to the expert scalar, not the category label.

Exact error decomposition. Adding and subtracting C_j and F_j yields

$$A_j - D_j = \underbrace{A_j - C_j}_{O_j = u'_j g_j} + \underbrace{C_j - F_j}_{N_j} - \underbrace{R_j}_{D_j - F_j}. \quad (3)$$

The three terms measure *reference-origin mismatch* O, *finite-step expert approximation* N, and *omitted routing response* R. The identity holds across routing support changes and does not require a differentiable router. Signed terms can cancel. Their absolute magnitudes are not additive error shares. This operational attribution of error is conditional on the chosen reference and frozen-route convention.

Controlled limits. Compensated translations of expert coordinates and references, with the routing function held fixed, preserve C while potentially changing A (Kindermans et al., 2017). Affine experts give $N = 0$, yet R may remain nonzero under continuously changing dense weights. Thus $C - D = -R$ in dense affine mixtures: a sign reversal can arise without expert nonlinearity or discrete selection. These two controls distinguish origin and routing effects without changing their response target.

Corrections address different targets. Centering removes O. Exact frozen-route evaluation removes N. Neither includes rerouting. In a smooth region, complete differentiation instead uses

$$S_j = (x_j - x'_j) \frac{\partial \ell}{\partial x_j}, \qquad \frac{\partial \ell}{\partial x_j} = \sum_e z_e \frac{\partial r_e(\psi(x))}{\partial x_j} + \sum_e r_e(\psi(x)) \frac{\partial z_e}{\partial x_j}. \quad (4)$$

For affine numerical encodings, $S - C$ is the displacement-scaled router derivative, motivating it as a priority for finite checks. S is still a local approximation. Exact endpoint evaluation is needed for D. Categories and hard support switches require finite evaluations. An endpoint bound $|R_j| \leq B_j = \frac{1}{2}\|z - z'\|_1 (\max_e r_e(u') - \min_e r_e(u'))$ also gives $|C_j - D_j| \leq |N_j| + B_j$. When $|F_j| > B_j$, it certifies that D_j and F_j have the same sign. It uses both endpoints and saves no queries.

Freezing the replacement route gives $F'_j = z'^{\top}(r - r')$, $R'_j = (z - z')^{\top} r$. Both conventions and their symmetric average are exact. We use the original route to match the score's computation and report convention sensitivity. Appendix B gives derivations and the ReLU endpoint qualification.

3 Related Work

IME combines instance-wise selection with interpretable/deep experts (Ismail et al., 2023), and InterpretCC selects user-defined feature groups (Swamy et al., 2025). MoE-X studies sparse expert activations, while cross-layer analyses examine routing contributions in language models (Yang et al., 2025; Li et al., 2026). Our controls instrument two routing families rather than reproduce these systems.

SENN distinguishes input-dependent coefficients from local predictor behavior (Alvarez-Melis & Jaakkola, 2018). Compensated translations expose gradient-times-input sensitivity (Kindermans et al., 2017), motivating our centering control. Additive models expose feature functions (Agarwal et al., 2021; Nori et al., 2019), while SHAP/TreeSHAP allocate feature contributions (Lundberg & Lee, 2017; Lundberg et al., 2020). Our ledger accounts for weighted expert logits rather than Shapley values over fields.

Removal-based explanations depend on the replacement, target, and aggregation (Covert et al., 2021). Infidelity measures attribution–response disagreement (Yeh et al., 2019). We trace it to computations within a mixture. ERASER measures evidence removal (DeYoung et al., 2020), whereas ROAR retrains after removal (Hooker et al., 2019). We query a fixed predictor. Feature dependence and out-of-distribution replacements remain concerns (Aas et al., 2021; Hase et al., 2021), even with matched budgets. Our endpoint response is neither a conditional Shapley value nor a causal effect. Probability calibration and correctness ranking are separate targets (Guo et al., 2017; Geifman & El-Yaniv, 2017).

4 Routed Models and Feature Scores

4.1 Separate Routing and Field-Aligned Expert Representations

MERN uses two encoders (Figure 1). The routing encoder $h(x)$ concatenates numerical projections and categorical embeddings, then applies an MLP and layer normalization. The expert encoder $u = \psi(x) \in \mathbb{R}^d$ retains one scalar per field: a standardized numerical value or a categorical embedding with a learned scalar projection. Its categorical parameters are separate. A field replacement changes only one expert coordinate, while routing can respond to all inputs.

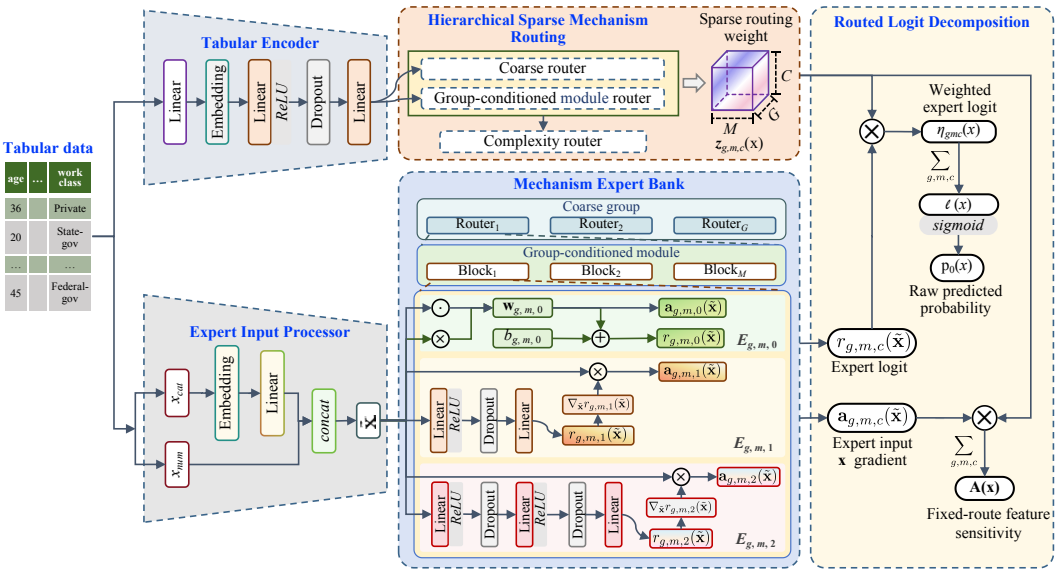

Figure 1: MERN architecture. Separate encoders feed routing and field-aligned affine experts ($c = 0$) or ReLU MLPs with one ($c = 1$) or two ($c = 2$) hidden layers. Weighted expert logits reconstruct ℓ and yield calibrated $p = \sigma(\ell/t)$. Feature scores A freeze routing. All experts receive all fields.

4.2 Hierarchical Sparse Routing and the Expert Ledger

Let TS_K denote softmax after retaining the largest K logits. The coarse and group-conditioned routers produce

$$\alpha(x) = \mathrm{TS}_2(W_g h(x) + b_g), \qquad \beta_g(x) = \mathrm{TS}_2(W_{m,g} q(x) + b_{m,g}), \tag{5}$$

$$q(x) = \tanh(W_s h(x) + b_s), \qquad \rho(x) = H(\alpha) + \sum_g \alpha_g H(\beta_g). \tag{6}$$

A learned map $\gamma(x) = \mathrm{softmax}(w_c \rho(x) + b_c)$ mixes three expert depths (Figure 2). These weights expose routing decisions, not semantic mechanisms or feature partitions. The depth map need not increase with routing entropy.

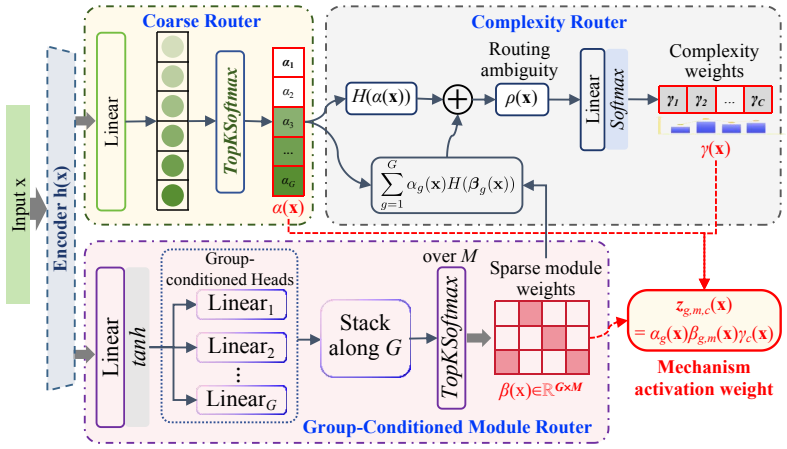

Figure 2: Routing factorization in MERN. Both sparse softmax operators retain $K = 2$. The depth weights $\gamma_0, \gamma_1, \gamma_2$ are shared across latent groups and blocks. The indices do not partition input fields or measure sample difficulty.

Each block (g, m) contains an affine expert $r_{g,m,0}(u) = w_{g,m,0}^\top u + b_{g,m,0}$ and one- and two-hidden-layer ReLU MLPs. Writing $e = (g, m, c)$,

$$z_e(x) = \alpha_g(x)\beta_{gm}(x)\gamma_c(x), \qquad \eta_e(x) = z_e(x) r_e(u), \tag{7}$$

$$\ell(x) = \sum_e \eta_e(x), \qquad p(x) = \sigma(\ell(x)/t). \tag{8}$$

The nonnegative weights sum to one, with at most $2 \times 2 \times 3 = 12$ active entries. All experts are evaluated before aggregation, so sparse weights do not imply computational savings. The signed ledger η_e includes expert biases and exactly reconstructs the logit; η_e/t reconstructs the temperature-scaled logit. Probabilities are not additive, and permuting expert indices leaves the predictor unchanged.

4.3 Local feature scores, descriptive scores, and training

The exported feature vector is $A_j = u_j \sum_e z_e \partial_j r_e(u)$: it uses $u_j w_{ej}$ for affine experts and local gradients for MLPs, holds routing fixed, and omits expert biases. Displays retain the five largest absolute scores. Section 2 specifies its relation to the finite response and the complete-gradient correction.

Auxiliary diagnostics include top-five concentration (TEC), route concentration, threshold margin, and algebraic subtraction of positive feature mass (d_{alg}). Appendix D defines them. Their correctness-ranking value is tested separately.

Training combines focal classification and pairwise ranking with router-balance/entropy, expert-depth, feature-concentration/sparsity, same-label stability, and negative-class penalties. Concentration and stability are directly encouraged, so favorable values of those diagnostics are not independent validation. The full loss, coefficients, and optimization settings remain in Appendix A.1. Validation selects the checkpoint, temperature t, and operating threshold. All subsequent replacement comparisons use the same fitted checkpoint and temperature.

5 Experimental Protocol

Predictor controls. The single-level control retains MERN's experts and active budget but changes routing parameters and regularization geometry, so it does not isolate topology. Independent affine

mixtures use a width-32 tanh router, dense or top-two softmax, BCE, and validation-NLL selection (maximum 60 epochs, patience 10), without auxiliary losses. Affine experts remove expert approximation error. Dense routing also removes support switching. Appendix A specifies the implementations.

Data and predictive baselines. We evaluate Zhongyuan, Taiwan, HELOC, Shoppers, Telco, Bank, Covertype, COMPAS, Adult, and MAGIC, using seeds 1–5 and stratified 64/16/20 fitting/validation/test splits (Becker & Kohavi, 1996; Moro et al., 2014; Bock, 2004). Covertype uses all 581,012 rows and the type-2-versus-rest task, with all 54 predictors verified against the official source (Blackard, 1998). Taiwan and Telco also match official source contents (Yeh, 2009; IBM, 2019). Zhongyuan's public provenance and feature-availability metadata remain incomplete. Appendices L–M document schemas, source checks, and protocol lineage.

Imputation, scaling, vocabularies, and references use fitting rows only. HELOC's special codes remain numerical, Shoppers' duration fields are omitted, and COMPAS's existing risk scores remain. Adult uses the 32,561-row source; Bank excludes call duration, and MAGIC contains simulated events. Validation is reused for model selection, temperature, threshold, and correctness estimators. Each dataset–seed shares 256 randomly selected test rows across the four routed predictors. A detected numerical failure prompted a stable, equivalent loss evaluation and uniform all-seed Taiwan reruns for both affected model families. Original attempts and the selection rule are documented in Appendix O.

References, selection, and aggregation. Each field is replaced by its fitting mean/mode or by the corresponding value in a fixed random fitting-row donor shared across methods. Partial donor replacements may violate dependencies. Neither reference is a causal intervention. For joint tests, predicted-class signed scores rank actually changed fields, selecting $\min(k, n_{\text{changed}})$ for $k = 1, 3, 5$. Negative scores remain eligible. We show $k = 3$ and recompute joint responses rather than sum field responses. A separate display protocol selects the top five absolute scores. Effect-size displays retain actually changed entries without a support filter, while a legacy sensitivity analysis additionally requires support for the predicted class. Random controls match counts alone or jointly match support, actual changes, and numerical/categorical counts.

Single-field outcomes are logit error, material-sign agreement (tolerance 10^{-6}), and within-observation rank agreement with D. Joint outcomes are signed probability changes $(2\hat{y} - 1)[p(x) - p(x')]$. Comparisons are within a predictor. Cross-model deletion magnitudes are not common quality units. We average observations within runs, datasets/models within seed for macros, then report five-seed means and sample SD. Overlapping holdouts are not independent replicates. The initial controlled three-seed cohort remains separate. All follow-ups are post hoc: their fixed protocols preceded the respective new computations, not the original test outcomes. All declared follow-up conditions are retained.

6 Results

6.1 Predictive performance and the role of the controls

Table 1 reports routed-model AUROC and CatBoost context (Prokhorenkova et al., 2018; Grinsztajn et al., 2022). CatBoost uses three seeds on HELOC/Shoppers/COMPAS and five elsewhere. EBM and local NAM-style controls remain in Appendix F. Their feature functions and TreeSHAP are not common-scale comparators for within-predictor response fidelity. Preprocessing and overlapping splits are shared, but objectives and tuning budgets differ. These comparisons establish neither predictive superiority nor negligible interpretability cost.

6.2 Centering helps, but leaves routing-related disagreement

Reference-dependent correction. For hierarchical MERN with donor references, centering reduces single-field logit MAE from .391 to .335, improves sign agreement from 63.5% to 75.3%, and increases joint probability drop from .192 to .244. Across all 200 runs, the donor $C - A$ gain has median 4.93 percentage points and range $-.39$–14.50 (Figure 3). Bank/hierarchical/seed 2 falls from .201 to .197, giving 199 improvements and one deterioration. Mean/mode references give 141

Table 1: Predictive context: AUROC, mean ± seed SD. Routed models use five seeds per dataset; CatBoost uses three on HELOC/Shoppers/COMPAS and five elsewhere. Overlapping seeds share splits and fitting-only preprocessing. Objectives and tuning budgets differ. AUPRC and EBM/NAM-style controls are in the appendix.

Dataset	Hierarchical	Single-level	Affine dense	Affine top-2	CatBoost
HELOC	0.783±0.008	0.785±0.006	0.785±0.008	0.784±0.007	0.797±0.005
Shoppers	0.912±0.009	0.912±0.003	0.913±0.014	0.914±0.006	0.927±0.005
COMPAS	0.747±0.011	0.744±0.017	0.746±0.014	0.742±0.012	0.763±0.007
Adult	0.910±0.003	0.909±0.004	0.913±0.003	0.912±0.003	0.927±0.002
Bank	0.787±0.007	0.791±0.006	0.797±0.011	0.796±0.006	0.804±0.011
MAGIC	0.930±0.004	0.929±0.005	0.924±0.002	0.922±0.003	0.936±0.002
Zhongyuan	0.646±0.013	0.650±0.019	0.639±0.022	0.629±0.018	0.679±0.013
Taiwan	0.771±0.008	0.772±0.005	0.771±0.009	0.769±0.008	0.779±0.007
Telco	0.840±0.010	0.841±0.010	0.841±0.008	0.839±0.009	0.845±0.011
Covertype	0.932±0.009	0.935±0.015	0.941±0.003	0.936±0.001	0.914±0.000

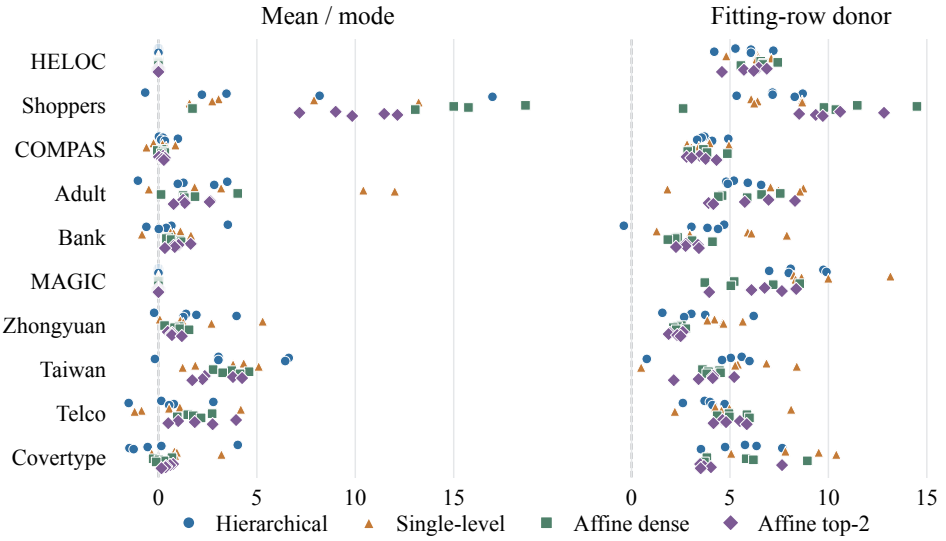

Figure 3: Centered minus original: joint probability drop (percentage points), $C - A$, at actual budget $k = 3$. Each panel shows 200 dataset–seed–predictor runs, with seeds offset vertically. Donors favor C in 199/200 runs, including one Bank reversal. Mean/mode gains are less uniform. These descriptive run means are not independent population replicates.

improving, 40 tied, and 19 worsening runs. Standardized means make $A \simeq C$ on numerical-only HELOC/MAGIC. Compensated categorical-coordinate translations preserve C within 9.7×10^{-7} in all 160 applicable runs. The 40 numerical-only runs are inapplicable.

Repeated and local donors. On the original 36-checkpoint cohort, five further fitting-row draws favor C in all 180 condition means. Nearest-ten fitting-row donors on HELOC favor it in all 60 conditions. Averaging draws precedes seed summaries. HELOC dense centered-display reversal incidence above a .05 probability change falls from 1.20% with random donors to .17% with local donors. Locality also changes replacement distances and changed-field counts. It does not establish conditional or causal validity. Appendix J retains the original-cohort sensitivity analyses.

Table 2: All-field centering and equal-budget selection over ten datasets and five seeds. Values average observations and then datasets within seed. MAE uses the same predictor's single-field D. Joint drop uses $\min(3, n_{\text{changed}})$ fields. Comparisons across predictors are not common attribution-quality units. Complete SDs, datasets, and budgets are released.

Reference	Predictor	MAE A	MAE C	Drop A	Drop C	Drop D
Mean/mode	Hierarchical	0.274	0.265	0.216	0.231	0.303
Mean/mode	Single-level	0.238	0.227	0.239	0.258	0.312
Mean/mode	Affine dense	0.304	0.281	0.190	0.212	0.267
Mean/mode	Affine top-2	0.290	0.266	0.199	0.218	0.257
Fit donor	Hierarchical	0.391	0.335	0.192	0.244	0.321
Fit donor	Single-level	0.365	0.296	0.213	0.273	0.333
Fit donor	Affine dense	0.450	0.346	0.185	0.236	0.284
Fit donor	Affine top-2	0.448	0.336	0.183	0.234	0.275

Individual Covertype indicator replacements can break one-hot validity. Replacing entire wilderness and soil blocks preserves it: centering gains 5.53 ± 2.07 percentage points at a three-group budget, improving 19/20 conditions (Appendix M.1). These budgets count groups, not fields, and one-hot validity does not imply ecological plausibility.

Routing remains relevant with affine experts. Across hierarchical and single-level dataset–seed–reference conditions, mean $|R| > |N|$ in 182/200. These absolute terms are not additive shares. In the affine mixtures, mean $|N| < 2.1 \times 10^{-7}$, yet dense routing reverses the sign between F and D in 18.1%/19.4% of changed-field comparisons under mean/mode and donor references (top-two: 16.0%/17.8%). Dense support never switches, so continuously changing weights suffice. These rates concern response direction, not prediction errors, and include small effects.

Complete-predictor gradient control. On numerical fields of all 200 checkpoints and both references, $S_j = (x_j - x'_j)\partial_j \ell(x)$ includes the router derivative, locally within the active branch. It improves sign agreement in 377/400 pairs, MAE in 325/400, and joint numerical $k = 3$ response in 261/400. The corresponding tie counts are 20, 18, and 58. Mean/mode MAE falls from .233 to .171, while donor MAE changes only from .318 to .313. Donor MAE deteriorates on Shoppers (.462 to .598) and Covertype (.855 to 1.095). Better local direction need not improve finite-step magnitude or joint selection. Of 4,680 predetermined numerical-gradient checks, 4,676 pass and four lie at piecewise boundaries (maximum passed error 2.80×10^{-8}). Categorical replacements are excluded. Appendix M retains every condition.

6.3 How large are the disagreements on displayed fields?

The 10^{-6} logit tolerance excludes numerical noise, not negligible probability effects. On numerical fields, we compare material C–D and S–D reversals jointly with $|p(x) - p(x^{(j)})| > \epsilon$. Both use the identical display of up to five largest absolute numerical C scores, intersected with actually changed fields, without a support filter; Telco has only three numerical fields. Fields below the effect threshold remain in the denominator. Thresholds $0, .001, .01, .05$ are descriptive and calibration-dependent, not validated utility criteria. The original all-field F–D sensitivity is retained separately in Appendix G. Full differentiation reduces large-effect direction disagreements but leaves a minority of displayed fields affected (Table 3). The release retains all thresholds, support strata, near-zero cases, and 32 empty donor displays. Observation-weighted rates differ from pooled counts when display sizes vary. Full counts are in Appendix M.

Prioritizing finite checks. We fix the numerical $|S|$ display of up to five fields and check up to two changed fields, chosen by $|S - C|$, $|S|$, or uniformly at random. Queried scores become exact D. Selection never sees D. Under donors, residual important reversals average 2.33% unchecked, 1.33% with random checks, 1.11% with magnitude checks, and .84% with routing-derivative priority. Mean/mode rates are 1.30%, .77%, .68%, and .47%. The priority beats random in 164/200 donor

Table 3: Numerical-field repair over 200 routed checkpoints: five-seed mean ± SD after averaging observations, four models, and ten datasets within seed. Reversal incidence requires opposite material signs and $|\Delta p| > 0.05$ on the identical changed numerical $|C|$ top-five display. Fields below the effect threshold remain in the denominator.

Reference	MAE C	MAE S	Reversal C (%)	Reversal S (%)
Mean/mode	0.233±0.016	0.171±0.012	2.468±0.207	1.443±0.151
Fit donor	0.318±0.013	0.313±0.024	3.808±0.186	2.560±0.233

conditions, ties in 14, and loses in 22. This tests allocation: queried errors disappear by construction. Appendix M reports seed variability, and the release retains all budgets 1/2/3/5 and exact random expectations. No wall-clock or human-benefit claim follows. Figure 4 illustrates the distinct repairs:

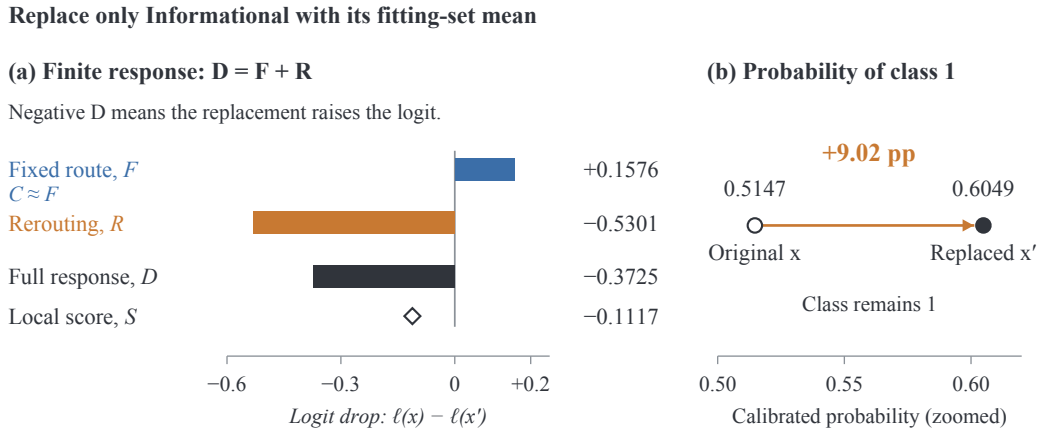

Figure 4: Shoppers, affine top-two, seed 1. Replacing Informational by its fitting mean gives $F > 0$ but $D = F + R < 0$, raising calibrated probability from .5147 to .6049. The local score S repairs direction, not magnitude, and is not an additive term. Active experts and predicted class remain unchanged. This post-hoc case is the lower median of 1,236 eligible high-impact reversals (Appendix G).

S corrects direction, and endpoint evaluation resolves the remaining magnitude error. Neither changes the original prediction or establishes human benefit.

Correctness is a separate target. Confidence provides an error-detection baseline (Hendrycks & Gimpel, 2017), evaluated by AURC (Geifman et al., 2019). On 18 hierarchical/single-level checkpoints (HELOC/Shoppers/COMPAS, seeds 1–3), auxiliary concentration and algebraic-subtraction diagnostics improve COMPAS AURC in 6/6 runs against a confidence-only fit, but only 2/6 against a fixed comparator using cubic confidence, predicted class, their interactions, and threshold margin. Small single-feature gains remain, but a stable combined gain is not established. These diagnostics differ from the decomposition terms. Appendix E retains all seven validation-only fits, coefficients, metrics, and adverse outcomes.

6.4 External transfer to SENN's native coefficients

We test the coefficient–response distinction in the original-author SENN identity-concept implementation (Alvarez-Melis & Jaakkola, 2018), where $\ell(x) = \sum_i \theta_i(x) h_i(x)$ and native explanations are coefficients θ, not finite responses. On HELOC, Adult, and MAGIC (five seeds), we compare numerical θ_j with the complete local derivative, and $C_j = (x_j - x'_j)\theta_j(x)$ and S_j with finite D_j. Both the released `grad3` arm and a non-official logit-aligned control use nonzero regularization and

Table 4: Original-author SENN and the non-official logit-aligned control, five seeds per arm (mean $\pm$ SD). Local agreement compares native θ_j with $\partial_j \ell$. Donor agreement compares derived $C_j = \Delta x_j \theta_j$ and $S_j = \Delta x_j \partial_j \ell$ with D_j. Sign agreement (%) averages observations over material numerical targets.

Dataset	Training arm	AUROC	Local θ sign	Donor C sign	Donor S sign
HELOC	Released	0.784 ± 0.006	83.41 ± 1.15	83.35 ± 1.04	96.17 ± 0.95
HELOC	Logit-aligned	0.783 ± 0.006	99.20 ± 0.38	99.14 ± 0.43	99.56 ± 0.25
Adult	Released	0.911 ± 0.002	87.54 ± 4.78	83.80 ± 3.14	91.97 ± 1.81
Adult	Logit-aligned	0.906 ± 0.002	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
MAGIC	Released	0.910 ± 0.001	82.62 ± 2.56	82.90 ± 2.19	95.95 ± 0.67
MAGIC	Logit-aligned	0.899 ± 0.006	98.62 ± 0.76	97.93 ± 0.83	98.60 ± 0.35

validation-NLL selection. The control jointly changes three regularizer details, so it does not isolate one cause. Appendix K documents source, forward checks, and target scales.

On released predictors, $C \to S$ improves donor sign agreement on all three datasets and lowers logit MAE from .090 to .035 (HELOC), .295 to .165 (Adult), and .351 to .196 (MAGIC). At $|\Delta p| > .05$, reversal incidence falls from .073% to .031%, 2.743% to 2.413%, and 1.141% to .766%. The logit-aligned control has no observed events at this threshold, but AUROC falls from .911 to .906 on Adult and .910 to .899 on MAGIC. It largely removes coefficient mismatch, not every finite-step error; this existing-model result neither refutes SENN's local principle nor establishes a cost-free correction.

7 Discussion and Conclusion

The decomposition separates reference-origin mismatch, expert approximation, and routing response, even when their signed effects cancel. Dense affine failures show why simple experts are insufficient. Residual errors after complete differentiation show why local corrections still need finite checks.

Implications and limits. Report the reference, frozen-route convention, and effect magnitude. Jointly reevaluate selected fields. The $|S - C|$ query priority is an empirical allocation rule, and native coefficients retain their own units. Exact accounting, response fidelity, and correctness information are distinct properties. These binary tabular experiments establish neither human benefit nor causal or domain-shift validity. SENN results concern its identity-concept implementation and our training control. Donor hybrids may violate dependencies. Local donors are not conditional samples. Validation and test reuse limit generalization. Leave-one/two-dataset analyses preserve positive average donor centering, sign, and joint gains but expose heterogeneous MAE reduction (Appendix N).

Reproducibility Statement

The supplement provides protocols, source/split hashes, per-run records, and offline reanalysis code. Source access limits and historical cohorts are documented. Fresh training requires source data and a new local code freeze. Raw data and checkpoints are excluded.

AI use statement

Generative AI tools assisted literature search, checking, and editing. The authors reviewed the assisted work, verified results and citations, and take responsibility for the final content.

REFERENCES

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021. doi: 10.1016/j.artint.2021.103502. URL https://doi.org/10.1016/j.artint.2021.103502.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4699–4711, 2021.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL https://doi.org/10.24432/C5XW20. Dataset.

Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. URL https://doi.org/10.24432/C50K5N. Dataset.

R. Bock. MAGIC gamma telescope. UCI Machine Learning Repository, 2004. URL https://doi.org/10.24432/C52C8B. Monte Carlo dataset.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://www.jmlr.org/papers/v22/20-1316.html.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.08206.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pp. 507–520, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1def1713ebf17722cbe300cfc1c88558-Abstract.html.

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.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

IBM. Telco customer churn sample data. IBM Cloud Pak for Data code-pattern repository, 2019. URL https://github.com/IBM/telco-customer-churn-on-icp4d/blob/1fd6fd70906479be1712e64cfc0ea89f58466629/data/Telco-Customer-Churn.csv. Pinned data-file version, September 30, 2019.

Aya Abdelsalam Ismail, Sercan Ö. Arik, Jinsung Yoon, Ankur Taly, Soheil Feizi, and Tomas Pfister. Interpretable mixture of experts. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=DdZoPUPm0a.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017. URL https://arxiv.org/abs/1711.00867.

Wengang Li, Lingqi Zhang, Toshio Endo, and Mohamed Wahib. Understanding cross-layer contributions to mixture-of-experts routing in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3efdd3cd3d70da09ec6ed3b3d17d2fd5-Abstract-Conference.html.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 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*, 2:56–67, 2020. doi: 10.1038/s42256-019-0138-9. URL https://www.nature.com/articles/s42256-019-0138-9.

Sérgio Moro, Paulo Rita, and Paulo Cortez. Bank marketing. UCI Machine Learning Repository, 2014. URL https://doi.org/10.24432/C5K306. Dataset; bank-additional-full version.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019. URL https://arxiv.org/abs/1909.09223.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Vinitra Swamy, Syrielle Montariol, Julian Blackwell, Jibril Frej, Martin Jaggi, and Tanja Käser. Intrinsic user-centric interpretability through global mixture of experts. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/47a09b2ea829d143862416e804ed7232-Abstract-Conference.html.

Xingyi Yang, Constantin Venhoff, Ashkan Khakzar, Christian Schroeder De Witt, Puneet K. Dokania, Adel Bibi, and Philip Torr. Mixture of experts made intrinsically interpretable. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71231–71248, 2025. URL https://proceedings.mlr.press/v267/yang25ag.html.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

I-Cheng Yeh. Default of credit card clients. UCI Machine Learning Repository, 2009. URL https://doi.org/10.24432/C55S3H. Dataset.

A Model, Data, and Optimization Details

Data and fitting-only preprocessing. This section records the original cohort and model definitions. The additional datasets and five-seed extension are specified separately in Appendix M; the tables below retain their labeled original three-seed scope. The original three tasks are HELOC bad credit performance (10,459 rows; 52.19% positive), Online Shoppers revenue (12,330; 15.47%), and COMPAS two-year recidivism (7,214; 45.07%). Their fixed numerical/categorical field counts are 23/0, 7/7, and 12/9. Seeds 1–3 use stratified 80/20 outer and 80/20 inner splits, giving approximately 64/16/20 fitting/validation/test proportions. Numerical medians, post-imputation means and population SDs, categorical modes, and vocabularies use only fitting rows. Neural numerical fields are standardized; native baselines receive imputed, unscaled numbers. Neural code zero is reserved for unknown categories and known categories start at one. Native inputs retain raw category strings without a train/test vocabulary union. All-missing numerical fields fall back to zero; zero scales become one. The manifest and `analysis/controlled_data.py` enumerate every retained field, source hash, split identity and transformation.

HELOC negative special codes remain numerical. Shoppers excludes `Administrative_Duration`, `Informational_Duration`, and `ProductRelated_Duration`. COMPAS retains existing risk scores and duplicated-name score/count fields; its exact 21-field schema is fixed in the released code. These retrospective processed tasks do not establish temporal availability, causal validity, fairness or deployment readiness.

Implementation dimensions. Section 4 defines the two encoders, router, expert bank, and exact logit ledger; Figure 2 shows the original routing diagram. Each categorical expert embedding has width $\min(8, \max(4, \lceil\sqrt{V_j}\rceil))$ before its learned scalar projection. The separate routing encoder uses its own embeddings and numerical projection, then a hidden MLP layer, an output projection, and layer normalization.

HELOC/Shoppers use routing width 72, encoder hidden width 96, and $G = M = 3$; COMPAS uses 96, 128, and $G = M = 4$. Expert hidden widths are 96 (one layer) and 128 (two layers), with dropout 0.08. The single-level control uses one coarse group and GM blocks, and increases block top-K from two to four, preserving the expert bank and maximum active count. This changes balance targets, entropy domains and effective penalty scales; inherited coefficients are not renormalized. It is an implementation control, not an isolated topology ablation.

Independent affine controls. Four affine experts share a fieldwise representation. Each categorical field has a learned scalar lookup initialized from $\mathcal{N}(0, 0.2^2)$; its unknown lookup stays zero. The independent router takes numerical fields and fixed one-hot categories through a width-32 tanh layer and four logits. Dense or top-two softmax mixes expert logits before the sigmoid; all experts are evaluated. Training uses unweighted BCE, AdamW (0.003 learning rate, 10^{-4} decay), batch size 256, at most 60 epochs, and minimum validation NLL with patience 10 and minimum improvement 10^{-7}. Affine initialization is PyTorch's default. There are no ranking, entropy, balance or explanation penalties. All seeds are retained; analytical frozen-route gradients and reloaded predictions are checked. These are local controls, not official IME/SENN reproductions.

A.1 Full Objective and Validation Rules

For a batch of B observations, define $p_{0i} = \sigma(\ell_i)$, $p_{ti} = y_i p_{0i} + (1-y_i)(1-p_{0i})$ and $\bar{c}_i = \sum_{c=0}^{2} c\gamma_{ic}$. The hierarchical/single-level objective is

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{focal}} + .01\mathcal{L}_{\text{bal}} + .001\mathcal{L}_{\text{ent}} + .01\mathcal{L}_{\text{cmp}} + .008\mathcal{L}_{\text{tec}} + .002\mathcal{L}_{\text{asp}} \\ &\quad + .004\mathcal{L}_{\text{stab}} + .004\mathcal{L}_{\text{unc}} + .75\mathcal{L}_{\text{rank}} + .03\mathcal{L}_{\text{neg}}, \\ \mathcal{L}_{\text{focal}} &= B^{-1}\sum_i (1-p_{ti})^{0.5}[-y_i \log p_{0i} - (1-y_i)\log(1-p_{0i})].\end{aligned} \tag{9}$$

There is no positive-class multiplier or monotonic constraint. Replacement sampling weights each fitting observation proportionally to $n_{y_i}^{-0.28}$. Let $\bar{\alpha}_g = B^{-1}\sum_i \alpha_{ig}$ and $\bar{\beta}_{gm} = B^{-1}\sum_i \beta_{igm}$. Then

$$\mathcal{L}_{\text{bal}} = G^{-1}\sum_g(\bar{\alpha}_g - 1/G)^2 + (GM)^{-1}\sum_{gm}(\bar{\beta}_{gm} - 1/M)^2,$$

$$\mathcal{L}_{\text{ent}} = B^{-1}\sum_i H(\alpha_i) + (BG)^{-1}\sum_{ig} H(\beta_{ig}),$$

$$v_i = \max\left(\rho_i / \max(\text{sg}(B^{-1}\sum_j \rho_j), \epsilon), 0\right), \quad s_i = v_i / \max(\text{sg}(\max_j v_j), \epsilon),$$

$$\mathcal{L}_{\text{cmp}} = B^{-1}\sum_i(\bar{c}_i - 2s_i)^2, \qquad \epsilon = 10^{-6}. \tag{10}$$

Here sg stops gradients; the numerator of s_i remains differentiable. Entropy regularization averages blocks uniformly, whereas ρ weights them by α. The normalization is batch-dependent.

The raw feature vector is $a_j = A_j = u_j \sum_e z_e \partial_j r_e(u)$; it holds routing fixed and omits biases and router derivatives. With A_5 denoting the five largest absolute entries,

$$\text{TEC} = \frac{\sum_{j\in A_5}|a_j|}{\max(\sum_j |a_j|, 10^{-6})}, \qquad \mathcal{L}_{\text{tec}} = B^{-1}\sum_i(1 - \text{TEC}_i), \tag{11}$$

$$\mathcal{L}_{\text{asp}} = \frac{(Bd)^{-1}\sum_{ij}|a_{ij}|}{B^{-1}\sum_i\sqrt{d^{-1}\sum_j a_{ij}^2 + 10^{-8}} + 10^{-6}}, \qquad \mathcal{L}_{\text{stab}} = |E_B|^{-1}\sum_{(i,j)\in E_B}\|\tilde{a}_i - \tilde{a}_j\|_2^2,$$

$$\mathcal{L}_{\text{unc}} = B^{-1}\sum_i[\bar{c}_i/2 - (1 - 2|\,\text{sg}(p_{0i}) - .5|)]^2.$$

For stability, E_B contains up to three nearest same-label neighbors per observation in Euclidean u-distance, excluding itself; distances, labels and neighbor choice are detached. $\tilde{a}_i = a_i / \max(\|a_i\|_2, 10^{-6})$. The term is zero with no valid pair. For within-batch positive/negative sets P_B, N_B,

$$\mathcal{L}_{\text{rank}} = \frac{\sum_{i\in P_B, j\in N_B}\log(1 + e^{-(\ell_i - \ell_j)})}{|P_B||N_B|}, \qquad \mathcal{L}_{\text{neg}} = |N_B|^{-1}\sum_{j\in N_B}[p_{0j} - .35]_+^2. \tag{12}$$

Terms requiring an absent class are zero. Concentration and stability are training preferences, not explanation-validity guarantees.

AdamW uses learning rate 4×10^{-4}, weight decay 5×10^{-5}, batch size 256, at most 40 epochs, and gradient-norm clipping at five. The checkpoint maximizes $.40\text{KS} + .30\text{F1} + .20\text{BAcc} + .10\text{AUROC}$ on validation, with threshold 0.5 for thresholded terms. The scheduler halves learning rate after two unimproved epochs; early stopping has patience 12 and restores the best checkpoint. Validation BCE fits temperature with LBFGS (learning rate .25, at most 50 iterations, $t \in [.2, 5]$). The operating threshold maximizes $.65\text{F1} + .35\text{BAcc}$ over 91 validation-score quantiles from 5% to 95%, then clips to $[.05, .95]$. Reusing validation for selection, calibration, thresholding and correctness-estimator fitting is disclosed; no separate calibration cohort is claimed.

B Derivations and Scope

Adding and subtracting C_j, F_j from $A_j - D_j$ gives Equation 3, regardless of router differentiability or support changes. The expert derivative is evaluated at the original input; ReLU uses its implementation-selected subgradient at zero. The fieldwise expert encoder is essential: replacing one field changes one expert coordinate. Vector-valued field encodings require block inner products; an encoder mixing fields does not satisfy this premise. A categorical reference is a second category's encoding, not a derivative of a discrete category.

For affine experts, $r_e(u) - r_e(u') = w_{ej}(u_j - u'_j)$ gives $F_j = C_j$. This also holds on an affine segment when the selected gradient agrees with its slope. A nondifferentiable endpoint needs care: for $r(u) = \text{ReLU}(u)$, $u = 0, u' = 1$, the usual derivative at zero yields $C = 0$ but $F = -1$.

Under $\tilde{u}_j = u_j + c$ and $\tilde{u}'_j = u'_j + c$, subtracting c times the incoming weight from every expert's first-layer bias preserves all activations, outputs and derivatives. The separate routing stream is unchanged, so C_j is invariant. A shared router would also require compensation. This statement covers the specified translation and consistently transformed reference, not arbitrary basis changes or nonlinear reparameterizations. Floating-point residuals are checked separately. For joint replacement S, define $C_S = \sum_{j \in S}(u_j - u'_j)g_j$ and evaluate F_S, R_S, D_S at the joint endpoint. Single-field finite responses need not sum to the joint response because experts and routing may interact.

Elementary endpoint bound. For $v = z - z'$, normalization implies $\sum_e v_e = 0$. Subtracting $m = (\max_e r'_e + \min_e r'_e)/2$ and applying the triangle inequality gives

$$|R_j| = \left| \sum_e v_e (r'_e - m) \right| \leq \tfrac{1}{2} \|z - z'\|_1 (\max_e r'_e - \min_e r'_e) =: B_j. \tag{13}$$

Thus $|F_j| > B_j$ guarantees that $D_j = F_j + R_j$ and F_j have the same sign; also $|C_j - D_j| \leq |N_j| + B_j$. The condition is sufficient, not necessary: uncovered cases are uncertified, not violations. It uses both evaluated endpoints, supplies no query saving, and is not a correctness probability or causal guarantee. Floating-point normalization slack and tolerances are recorded in the exports.

Endpoint convention. The original convention uses $F_0 = z^\top (r - r')$, $R_0 = (z - z')^\top r'$; the replacement convention uses $F_1 = z'^\top (r - r')$, $R_1 = (z - z')^\top r$; the symmetric convention averages them. For $z = (.8, .2)$, $z' = (.2, .8)$, $r = (2, 0)$, $r' = (0, 1)$, the three (F, R) pairs are $(1.4, -.6)$, $(-.4, 1.2)$ and $(.5, .3)$, all reconstructing $D = .8$. Opposition and sign reversal therefore depend on which route is frozen. We abbreviate $F = F_0$ and use $R^* = F - D = -R_0$ in discrepancy tables.

C Original Three-Seed Mechanism Comparisons

Tables below retain all three datasets, three seeds, four predictors, both references and all declared budgets. We average fields within observations when applicable, observations within runs, and datasets within each seed for macro results. Parentheses denote sample SD across three seeds, not standard errors or independent-population confidence intervals. References use fitting rows only; donor references are whole fitting rows shared across methods, although partial replacement can form atypical hybrids. The 256 audit observations and source-row identities agree across models. Effects below 10^{-6} are nonmaterial; actual change is inequality of encoded input values, not a threshold on prediction response.

Support and no-op checks. In HELOC/hierarchical/seed 3 under mean/mode replacement, exact positive-weight support changes on 12.8% of changed-field interventions, but support above 10^{-6} changes on 0.0% and material rerouting occurs on 0.0%. An exact switch of extremely small weights need not have a consequential response. Changed/no-op field counts are identical across predictors: mean/mode gives $16.11(.05)/3.22(.05)$; donor gives $13.13(.15)/6.20(.15)$ (macro mean(seed SD)). Raw A is material on nearly all no-op fields, conditional on having no-ops. No-ops receive no explanatory credit; undefined groups remain missing.

The original display-then-support rule is retained as a sensitivity analysis. Its count-only random controls and controls jointly matching support eligibility, actual changes and field type are saved with every mask. This separate rule can select different counts across methods and is not substituted for the primary equal-actual-budget test. Random masks are averaged within observations, never counted as independent training runs. Per-dataset support, endpoint, timing, sign agreement, and all three pairwise endpoint disagreements remain in the CSVs.

Table 5: Every dataset, predictor and reference: logit MAE and within-observation absolute-rank correlation with the same predictor's D. Cells are mean(SD) over three seeds. M: mean/mode; D: fitting donor; H: hierarchical; S: single-level; AD/AT: affine dense/top-two. Only actually changed fields enter these metrics. Constant rank vectors are missing, not zero. Sign agreement, RMSE, signed ranks and valid counts remain in the CSVs.

Ref.	Dataset	Model	MAE A	MAE C	MAE F	ρ_A	ρ_C	ρ_F
M	HELOC	H	.057(.040)	.057(.040)	.049(.043)	.830(.118)	.830(.118)	.878(.111)
M	HELOC	S	.069(.014)	.069(.014)	.065(.013)	.791(.023)	.791(.023)	.820(.027)
M	HELOC	AD	.056(.015)	.056(.015)	.056(.015)	.867(.019)	.867(.019)	.867(.019)
M	HELOC	AT	.037(.002)	.037(.002)	.037(.002)	.899(.010)	.899(.010)	.899(.010)
M	Shoppers	H	.467(.055)	.457(.051)	.443(.056)	.547(.019)	.580(.050)	.614(.051)
M	Shoppers	S	.432(.061)	.416(.080)	.404(.075)	.531(.068)	.564(.086)	.582(.073)
M	Shoppers	AD	.589(.031)	.498(.038)	.498(.038)	.400(.135)	.532(.097)	.532(.097)
M	Shoppers	AT	.496(.093)	.398(.064)	.398(.064)	.388(.086)	.559(.067)	.559(.067)
M	COMPAS	H	.061(.019)	.059(.017)	.058(.018)	.731(.019)	.769(.045)	.776(.047)
M	COMPAS	S	.048(.043)	.046(.041)	.045(.041)	.692(.137)	.732(.143)	.740(.155)
M	COMPAS	AD	.035(.009)	.034(.008)	.034(.008)	.742(.031)	.785(.048)	.785(.048)
M	COMPAS	AT	.025(.020)	.024(.019)	.024(.019)	.822(.093)	.857(.086)	.857(.086)
D	HELOC	H	.161(.041)	.104(.065)	.089(.079)	.478(.123)	.762(.149)	.824(.158)
D	HELOC	S	.169(.027)	.126(.026)	.121(.026)	.424(.018)	.700(.043)	.732(.046)
D	HELOC	AD	.185(.021)	.103(.025)	.103(.025)	.521(.037)	.808(.030)	.808(.030)
D	HELOC	AT	.157(.008)	.072(.005)	.072(.005)	.533(.017)	.847(.011)	.847(.011)
D	Shoppers	H	.549(.056)	.470(.043)	.460(.062)	.285(.008)	.544(.008)	.574(.015)
D	Shoppers	S	.490(.101)	.431(.103)	.404(.104)	.259(.097)	.512(.120)	.553(.117)
D	Shoppers	AD	.599(.053)	.457(.057)	.457(.057)	.187(.097)	.555(.021)	.555(.021)
D	Shoppers	AT	.536(.036)	.377(.027)	.377(.027)	.222(.031)	.626(.030)	.626(.030)
D	COMPAS	H	.128(.021)	.097(.022)	.097(.022)	.464(.079)	.715(.074)	.718(.076)
D	COMPAS	S	.098(.073)	.074(.062)	.074(.063)	.454(.125)	.699(.170)	.704(.188)
D	COMPAS	AD	.128(.007)	.058(.016)	.058(.016)	.503(.080)	.745(.078)	.745(.078)
D	COMPAS	AT	.115(.027)	.043(.033)	.043(.033)	.592(.131)	.839(.118)	.839(.118)

Table 6: All prespecified joint-selection budgets. Drop is the signed predicted-class probability change after jointly replacing fields ranked by A, C, F, D. Negative-ranked fields remain eligible. Excess C subtracts ten random masks matched on actual and field-type counts. Cells are three-dataset macro mean(seed SD). Each seed first averages datasets. Count is the actual number of changed fields. These are within-predictor comparisons, not a common attribution scale across models.

Ref.	Model	k	Drop A	Drop C	Drop F	Drop D	Excess C	Count
M	H	1	.076(.004)	.082(.006)	.094(.014)	.184(.016)	.065(.005)	1.00(.00)
M	H	3	.207(.030)	.213(.033)	.224(.034)	.276(.029)	.156(.025)	3.00(.00)
M	H	5	.289(.032)	.291(.030)	.295(.031)	.313(.026)	.195(.018)	5.00(.00)
M	S	1	.074(.013)	.086(.003)	.101(.007)	.187(.018)	.069(.003)	1.00(.00)
M	S	3	.205(.033)	.221(.029)	.230(.030)	.285(.025)	.166(.022)	3.00(.00)
M	S	5	.276(.042)	.286(.035)	.288(.037)	.321(.030)	.191(.019)	5.00(.00)
M	AD	1	.071(.008)	.115(.026)	.115(.026)	.172(.007)	.093(.025)	1.00(.00)
M	AD	3	.171(.019)	.227(.023)	.227(.023)	.256(.006)	.168(.022)	3.00(.00)
M	AD	5	.246(.020)	.280(.004)	.280(.004)	.290(.005)	.186(.008)	5.00(.00)
M	AT	1	.072(.006)	.094(.022)	.094(.022)	.155(.023)	.075(.018)	1.00(.00)
M	AT	3	.164(.027)	.196(.035)	.196(.035)	.235(.023)	.142(.030)	3.00(.00)
M	AT	5	.221(.034)	.260(.026)	.260(.026)	.271(.021)	.172(.021)	5.00(.00)
D	H	1	.077(.007)	.122(.008)	.124(.013)	.177(.007)	.095(.004)	1.00(.00)
D	H	3	.164(.013)	.224(.016)	.229(.018)	.268(.009)	.155(.008)	3.00(.00)
D	H	5	.215(.016)	.265(.019)	.267(.021)	.294(.011)	.162(.010)	4.99(.00)
D	S	1	.079(.004)	.120(.008)	.128(.009)	.177(.007)	.095(.007)	1.00(.00)
D	S	3	.162(.007)	.218(.009)	.225(.013)	.271(.014)	.151(.002)	3.00(.00)
D	S	5	.212(.012)	.260(.017)	.264(.019)	.295(.019)	.159(.005)	4.99(.00)
D	AD	1	.079(.008)	.120(.010)	.120(.010)	.161(.002)	.097(.008)	1.00(.00)
D	AD	3	.158(.009)	.231(.005)	.231(.005)	.254(.002)	.165(.004)	3.00(.00)
D	AD	5	.209(.008)	.273(.003)	.273(.003)	.285(.004)	.175(.001)	4.99(.00)
D	AT	1	.080(.005)	.118(.008)	.118(.008)	.156(.006)	.094(.006)	1.00(.00)
D	AT	3	.156(.006)	.223(.005)	.223(.005)	.242(.009)	.158(.005)	3.00(.00)
D	AT	5	.203(.015)	.265(.011)	.265(.011)	.272(.009)	.163(.003)	4.99(.00)

Table 7: Discrepancy magnitudes and support checks, macro mean(seed SD). $O = A - C$, $N = C - F$, $R^* = F - D$. Their absolute means are not additive shares. Switch is exact nonzero support change. Stable R is material rerouting conditional on unchanged support. Cert. is coverage of $|F| > B$. Tightness is $|R|/B$ where $B > 10^{-6}$. All endpoint, centered-error and sign-bound violation counts are zero. Both endpoints are already evaluated.

Ref.	Model	$\lvert O\rvert$	$\lvert N\rvert$	$\lvert R^*\rvert$	Switch%	Stable R%	Cert.%	Tightness
M	H	.015(.004)	.022(.003)	.183(.027)	18.7(2.5)	88.9(19.2)	25.8(13.4)	.246(.085)
M	S	.015(.005)	.021(.009)	.171(.040)	19.6(1.8)	100.0(.0)	23.8(8.0)	.310(.037)
M	AD	.050(.003)	.000(.000)	.196(.011)	.0(.0)	100.0(.0)	47.1(6.0)	.591(.029)
M	AT	.046(.007)	.000(.000)	.153(.026)	8.0(1.3)	99.7(.4)	59.3(2.2)	.558(.055)
D	H	.107(.011)	.060(.014)	.216(.040)	25.3(3.3)	88.9(19.2)	25.4(14.0)	.266(.053)
D	S	.095(.016)	.055(.022)	.200(.060)	27.5(2.4)	99.9(.2)	21.4(6.8)	.305(.036)
D	AD	.163(.003)	.000(.000)	.206(.012)	.0(.0)	100.0(.0)	45.6(4.1)	.587(.032)
D	AT	.157(.012)	.000(.000)	.164(.012)	12.3(1.7)	99.7(.4)	56.0(1.9)	.556(.055)

Table 8: Endpoint-convention sensitivity (percent, macro mean(seed SD)). Reversal means material $FD < 0$. Harm changes positive predicted-class F to negative D. Rescue is the reverse, without implying a change in prediction correctness. Disagree compares original/replacement reversal indicators on identical interventions. Similar aggregate rates can describe different events.

Ref.	Model	Orig.	Repl.	Sym.	Harm orig.	Rescue orig.	Disagree
M	H	14.2(2.5)	14.7(2.5)	14.4(2.5)	8.4(1.1)	5.7(1.5)	1.8(.1)
M	S	17.2(2.5)	17.3(2.4)	17.2(2.4)	9.9(.5)	7.3(2.1)	1.2(.3)
M	AD	16.4(1.4)	16.3(1.2)	16.3(1.3)	7.6(1.1)	8.8(1.0)	1.8(.0)
M	AT	11.6(2.5)	11.6(2.2)	11.5(2.3)	5.9(.8)	5.6(2.2)	1.8(.6)
D	H	16.1(3.4)	15.9(3.5)	15.8(3.4)	7.6(1.5)	8.5(1.9)	2.9(.4)
D	S	18.6(2.0)	18.6(2.2)	18.6(2.0)	9.3(.7)	9.4(1.3)	1.8(.6)
D	AD	17.9(1.5)	17.8(1.4)	17.8(1.5)	9.5(.7)	8.4(.8)	3.2(.0)
D	AT	13.9(2.9)	14.0(2.7)	13.8(2.8)	7.3(1.5)	6.6(1.5)	2.8(1.0)

Table 9: Descriptive synchronized cost: seconds, macro mean(seed SD). Base is one original forward/gradient shared by A, C, shown twice but incurred once. Ref. encodes the reference. Fields evaluates endpoints shared by F, D. Joint includes cached selected/random-mask forwards for both selection rules and all budgets. Row counts are evaluated observations, not independent units. Different warm-up/batching costs prevent a speedup claim.

Ref.	Model	Base s	Ref. s	Fields s	Joint s	Field rows	New joint rows
M	H	.11(.12)	.001(.000)	.16(.01)	1.87(.13)	4949(0)	50097(986)
M	S	.11(.08)	.001(.000)	.15(.00)	1.71(.02)	4949(0)	50315(974)
M	AD	.02(.04)	.001(.000)	.01(.00)	.26(.02)	4949(0)	47352(1229)
M	AT	.03(.04)	.001(.000)	.01(.00)	.28(.02)	4949(0)	48174(754)
D	H	.11(.12)	.005(.001)	.18(.02)	1.71(.16)	4949(0)	42681(623)
D	S	.11(.08)	.005(.001)	.15(.01)	1.42(.02)	4949(0)	42309(504)
D	AD	.02(.04)	.004(.001)	.01(.00)	.24(.01)	4949(0)	41159(732)
D	AT	.03(.04)	.022(.001)	.01(.00)	.25(.01)	4949(0)	41126(242)

Table 10: Compensated categorical-coordinate translations: maximum centered-score error and raw-score change over three seeds, both references and shifts $+1, -1$. All 24 applicable runs pass. The 12 numeric-only HELOC runs are explicitly skipped. Logit, ledger and routing errors also satisfy 10^{-4}. These maxima are numerical checks, not performance averages.

Dataset	Model	Max centered error	Max raw change
Shoppers	H	2.48×10^{-7}	1.40×10^{0}
Shoppers	S	1.13×10^{-7}	1.04×10^{0}
Shoppers	AD	3.07×10^{-7}	1.25×10^{0}
Shoppers	AT	2.60×10^{-7}	1.41×10^{0}
COMPAS	H	6.10×10^{-9}	1.49×10^{-1}
COMPAS	S	0.00×10^{0}	1.98×10^{-1}
COMPAS	AD	1.82×10^{-7}	4.09×10^{-1}
COMPAS	AT	2.14×10^{-7}	3.66×10^{-1}

D Original Correctness Estimators

All 18 hierarchical/single-level checkpoints use their complete test splits to assess correctness at the saved threshold. Confidence is $\hat{y}p + (1-\hat{y})(1-p)$, which differs from $\max(p, 1-p)$ when the threshold is not .5. With s positive coarse weights, $q_r = 1 - H(\alpha)/\log s$, taking value one if $s \le 1$; joint concentration applies the same rule to z. In the single-level model $q_r = 1$. For the five largest absolute raw terms,

$$d_{\text{alg}} = \sigma(\ell) - \sigma(\ell - \sum_{j \in A_5} [a_j]_+), \qquad q_p = |p - \tau| / \max(\tau, 1-\tau), \tag{14}$$

$$q_e = .65\,\text{TEC} + .35\min(d_{\text{alg}}/.25, 1), \qquad T_{\text{proxy}} = .45q_p + .35q_e + .20q_r. \tag{15}$$

The proxy uses uncalibrated logits and positive feature mass even for negative-class decisions, performs no replacement, and is nonnegative by construction. The fixed T mixes a calibrated margin with that uncalibrated proxy; its weights do not learn correctness. Definitions are checked against the original forward/wrapper formulas.

Meta-C uses confidence; Meta-C+coarse adds q_r; Meta-C+route adds coarse and joint concentration; Meta-C+feature adds TEC and d_{alg}; Meta-All adds all four diagnostics. Each uses validation-only standardization and L2 logistic regression with $C = 1$, intercept, LBFGS, 2,000-iteration cap, tolerance 10^{-8} and no class weighting or search. Validation has already served checkpoint/calibration/threshold selection. No arm is selected by test results. All 18 Meta-C slopes are positive; AUROC and AURC match confidence within 10^{-10}, so this arm recalibrates correctness probabilities without changing the observed ranking.

For decreasing-score order, $R_k = k^{-1}\sum_{i \le k} \mathbb{1}[\hat{y}_i \ne y_i]$ and $\text{AURC} = n^{-1}\sum_{k=1}^{n} R_k$, averaging exact score ties analytically. Correctness Brier targets correctness rather than the positive class. For raw heuristics it only describes squared error when their numerical values are treated as probabilities. Source/label hashes, split identities, fitted coefficients, validation standardization and score formulas are checked before aggregation.

The original five-arm AUROC, AURC, Brier and paired tables are retained with all seed records in the anonymous numerical supplement. The stronger follow-up below supersedes the one-dimensional confidence comparison for the incremental-information question.

E Stronger controls for correctness ranking

This reviewer-motivated follow-up retains all 18 hierarchical/single-level checkpoints, their saved probabilities, temperatures, operating thresholds, and complete validation/test splits. Write $c = \hat{y}p + (1-\hat{y})(1-p)$ and $m = |p-\tau|/\max(\tau, 1-\tau)$. Class-C uses $(c, \hat{y}, c\hat{y})$; Margin-C additionally uses m. The fixed Strong-C design is

$$(c, c^2, c^3, \hat{y}, c\hat{y}, c^2\hat{y}, c^3\hat{y}, m). \tag{16}$$

These are functions of the predictor probability and fixed threshold only. Since $m = [c + \tau - 1 + \hat{y}(1-2\tau)]/\max(\tau, 1-\tau)$, margin is redundant and changes the L2 parameterization rather than adding information. Each column is standardized on the reused validation partition. Every arm uses the same logistic estimator (L2, $C = 1$, intercept, `lbfgs`, maximum 2000 iterations, tolerance 10^{-8}); no degree, feature or regularization setting is selected on test outcomes.

Three fixed additions to Strong-C use TEC, d_{alg}, or both. These retain the earlier concentration and algebraic subtraction definitions; they do not evaluate the new intervention terms C, N, R. The primary paired contrast is Strong-C+both minus Strong-C. All three endpoints are reported: AURC and correctness Brier (lower is better), and correctness AUROC (higher is better). Single-diagnostic additions are secondary; the original Meta-F is an unchanged contextual arm. Tables report all fixed arms and all dataset/router groups. Seed-level metrics, fitted coefficients, validation scaling and paired differences are in the accompanying records.

The combined COMPAS advantage over Meta-C does not persist consistently: only 2/6 runs improve over Strong-C (mean paired AURC difference -0.000091, median $+0.000322$). Small architecture-specific secondary increments remain: TEC alone improves hierarchical COMPAS in 3/3 runs (mean -0.001226), and d_{alg} alone improves single-level COMPAS in 3/3 (mean -0.000561). We do not

select a winning arm from these results. Nor does Strong-C uniformly dominate earlier diagnostic combinations: the original Meta-F retains lower mean AURC on hierarchical COMPAS. Thus estimator parameterization matters, and neither universal diagnostic utility nor absence of diagnostic information follows.

This follow-up was motivated by previous test findings. The validation partition was already used for predictor selection, calibration and original meta fitting; it is not an independent calibration cohort. Shared holdouts across seeds are not independent population replications. Independent reconstruction verified 453,726 saved meta scores (maximum absolute discrepancy 8.8×10^{-15}), all metrics (maximum 6.2×10^{-16}), validation-only scaling, disjoint validation/test rows and unchanged numeric input hashes.

Table 11: Primary stronger-confidence contrast. Absolute AURC and paired differences are mean $\pm$ sample SD across three seeds. Median differences and improved-seed counts expose run heterogeneity. Negative differences favor adding both diagnostics.

Dataset	Router	Strong-C AURC	+both AURC	Paired Δ	Median Δ	Better
HELOC	H	0.1767 ± 0.0070	0.1761 ± 0.0079	-0.0006 ± 0.0011	-0.0005	2/3
HELOC	S	0.1775 ± 0.0077	0.1784 ± 0.0085	$+0.0008 \pm 0.0008$	$+0.0005$	0/3
Shoppers	H	0.0265 ± 0.0034	0.0265 ± 0.0033	$+0.0000 \pm 0.0001$	-0.0000	2/3
Shoppers	S	0.0250 ± 0.0011	0.0258 ± 0.0019	$+0.0008 \pm 0.0014$	$+0.0001$	1/3
COMPAS	H	0.2268 ± 0.0153	0.2268 ± 0.0141	$+0.0000 \pm 0.0012$	$+0.0006$	1/3
COMPAS	S	0.2387 ± 0.0277	0.2386 ± 0.0280	-0.0002 ± 0.0009	$+0.0002$	1/3

Table 12: All fixed correctness-ranking arms: AURC (lower is better). Entries are mean(sample SD) across three seeds, with leading zeros omitted. H: hierarchical; S: single-level. The three rows after Strong-C add the indicated original diagnostic(s).

	HELOC		Shoppers		COMPAS	
Arm	H	S	H	S	H	S
Meta-C	.1768(.0070)	.1778(.0090)	.0274(.0037)	.0261(.0014)	.2293(.0211)	.2495(.0316)
Class-C	.1766(.0072)	.1773(.0080)	.0268(.0032)	.0252(.0011)	.2277(.0168)	.2395(.0301)
Margin-C	.1766(.0072)	.1772(.0081)	.0268(.0033)	.0252(.0011)	.2278(.0169)	.2395(.0301)
Strong-C	.1767(.0070)	.1775(.0077)	.0265(.0034)	.0250(.0011)	.2268(.0153)	.2387(.0277)
Strong-C+TEC	.1764(.0073)	.1782(.0085)	.0263(.0037)	.0254(.0019)	.2256(.0152)	.2388(.0282)
Strong-C+d_{alg}	.1763(.0079)	.1776(.0076)	.0266(.0034)	.0250(.0012)	.2278(.0136)	.2382(.0277)
Strong-C+both	.1761(.0079)	.1784(.0085)	.0265(.0033)	.0258(.0019)	.2268(.0141)	.2386(.0280)
Old Meta-F	.1770(.0067)	.1788(.0087)	.0277(.0028)	.0276(.0017)	.2241(.0194)	.2396(.0301)

Table 13: All fixed arms: correctness AUROC (higher is better). Entries are mean(sample SD), with the same seeds and estimators as Table 12.

	HELOC		Shoppers		COMPAS	
Arm	H	S	H	S	H	S
Meta-C	.6824(.0170)	.6830(.0223)	.8734(.0034)	.8731(.0047)	.6493(.0282)	.6634(.0276)
Class-C	.6832(.0162)	.6847(.0206)	.8787(.0028)	.8808(.0020)	.6521(.0260)	.6819(.0148)
Margin-C	.6833(.0161)	.6848(.0204)	.8786(.0028)	.8808(.0020)	.6521(.0259)	.6820(.0148)
Strong-C	.6826(.0171)	.6838(.0221)	.8811(.0031)	.8824(.0015)	.6520(.0256)	.6818(.0152)
Strong-C+TEC	.6828(.0138)	.6819(.0234)	.8807(.0044)	.8809(.0028)	.6530(.0248)	.6817(.0151)
Strong-C+d_{alg}	.6827(.0143)	.6834(.0215)	.8807(.0045)	.8823(.0013)	.6496(.0273)	.6828(.0159)
Strong-C+both	.6829(.0128)	.6807(.0235)	.8805(.0047)	.8803(.0030)	.6505(.0262)	.6823(.0156)
Old Meta-F	.6815(.0140)	.6820(.0229)	.8739(.0043)	.8734(.0013)	.6543(.0279)	.6789(.0157)

Table 14: All fixed arms: Brier against correctness (lower is better). Entries are mean(sample SD), with the same seeds and estimators as Table 12.

	HELOC		Shoppers		COMPAS	
Arm	H	S	H	S	H	S
Meta-C	.1896(.0031)	.1899(.0018)	.0812(.0025)	.0805(.0035)	.2062(.0045)	.2116(.0094)
Class-C	.1895(.0032)	.1895(.0022)	.0768(.0026)	.0760(.0018)	.2054(.0032)	.2074(.0051)
Margin-C	.1895(.0032)	.1895(.0022)	.0769(.0027)	.0760(.0018)	.2054(.0032)	.2074(.0051)
Strong-C	.1896(.0031)	.1897(.0025)	.0762(.0027)	.0756(.0019)	.2061(.0037)	.2084(.0061)
Strong-C+TEC	.1895(.0034)	.1898(.0025)	.0763(.0026)	.0757(.0018)	.2059(.0036)	.2083(.0062)
Strong-C+d_{alg}	.1895(.0036)	.1897(.0025)	.0763(.0025)	.0755(.0021)	.2062(.0031)	.2081(.0059)
Strong-C+both	.1895(.0037)	.1899(.0024)	.0764(.0024)	.0756(.0019)	.2060(.0032)	.2081(.0059)
Old Meta-F	.1897(.0033)	.1899(.0017)	.0781(.0018)	.0774(.0026)	.2052(.0036)	.2087(.0064)

F Original Controlled Comparisons and Predictive Context

These five-model comparisons use the same fitting-only data, seeds 1–3, splits, calibration, threshold selection and 256 audit rows. EBM has main effects only, eight outer bags and at most 5,000 rounds. CatBoost uses at most 300 depth-six trees, learning rate .05 and validation early stopping; its field scores are TreeSHAP terms. The local NAM-style additive NN has one independent 64–32 ReLU network per field, categorical embeddings, a global bias and fitting-mean-centered terms. It uses unweighted BCE, AdamW (learning rate 4×10^{-4}, decay 5×10^{-5}), batch size 256, at most 40 epochs, validation-NLL selection and patience 12. Each baseline's random state is its run seed. Shared evaluation does not imply matched losses/compute or comprehensive tuning.

The original replacement rule keeps predicted-class-supporting fields among the five largest absolute terms and replaces them jointly with fitting means/modes. Empty sets remain in every primary average. Signed drop is $(2\hat{y} - 1)(p - p^{-S})$, without clipping; ten random masks match the selected count. The occlusion comparator ranks individual field responses at the same count. These controls establish behavior under the specified reference, not causal importance or comparisons on a common scale across different predictors.

Table 15: Complete predictive context: the original five-model comparison (45 trained runs) and independent affine controls (18 runs), mean(seed SD). AUPRC is average precision. Every model receives validation temperature scaling and the common threshold rule. Time includes training and initial evaluation under concurrent execution and is not hardware-normalized. Predictive losses and compute budgets differ.

Dataset	Model	AUROC	AUPRC	Brier	NLL	BAcc	Seconds
HELOC	H	.788(.005)	.786(.010)	.188(.002)	.557(.005)	.711(.011)	62.4(7.0)
HELOC	S	.788(.004)	.786(.008)	.188(.002)	.558(.006)	.710(.011)	68.1(17.8)
HELOC	EBM	.797(.002)	.790(.004)	.184(.001)	.548(.003)	.717(.011)	6.4(.1)
HELOC	Additive NN	.793(.006)	.786(.007)	.185(.003)	.552(.007)	.714(.015)	33.0(1.5)
HELOC	CatBoost	.797(.005)	.791(.002)	.184(.002)	.547(.005)	.718(.009)	3.9(.2)
HELOC	AD	.787(.005)	.780(.003)	.188(.002)	.559(.006)	.709(.010)	6.3(.6)
HELOC	AT	.785(.003)	.781(.001)	.189(.001)	.561(.003)	.711(.007)	4.4(.8)
Shoppers	H	.909(.009)	.711(.025)	.081(.003)	.263(.012)	.819(.009)	68.9(10.3)
Shoppers	S	.913(.002)	.710(.015)	.080(.003)	.261(.006)	.821(.007)	70.6(12.7)
Shoppers	EBM	.914(.005)	.691(.012)	.076(.002)	.251(.006)	.820(.009)	6.3(.4)
Shoppers	Additive NN	.910(.003)	.675(.014)	.078(.002)	.257(.004)	.825(.010)	28.1(.8)
Shoppers	CatBoost	.927(.005)	.739(.023)	.072(.003)	.233(.009)	.830(.006)	7.6(.4)
Shoppers	AD	.918(.003)	.722(.021)	.075(.003)	.244(.006)	.824(.003)	10.4(1.4)
Shoppers	AT	.914(.003)	.714(.022)	.076(.003)	.249(.005)	.818(.001)	9.4(2.2)
COMPAS	H	.743(.013)	.694(.014)	.206(.006)	.602(.013)	.679(.013)	65.9(37.0)
COMPAS	S	.735(.017)	.689(.019)	.211(.009)	.611(.020)	.654(.028)	58.6(30.8)
COMPAS	EBM	.758(.004)	.729(.006)	.198(.001)	.580(.001)	.688(.003)	6.0(.4)
COMPAS	Additive NN	.756(.007)	.721(.007)	.199(.002)	.584(.005)	.681(.005)	19.5(3.9)
COMPAS	CatBoost	.763(.007)	.732(.005)	.196(.002)	.577(.005)	.692(.012)	5.5(.7)
COMPAS	AD	.740(.010)	.700(.018)	.205(.003)	.598(.009)	.674(.011)	4.1(.3)
COMPAS	AT	.737(.008)	.692(.017)	.207(.003)	.602(.008)	.669(.014)	3.5(.9)

The original joint-selection finite-change decomposition reconstructs all hierarchical/single-level logit changes within 9.54×10^{-7}. Compensated categorical shifts preserve logits/ledgers within the same tolerance yet change the hierarchical leading feature on $20.7(5.1)\%$ of Shoppers and $6.6(9.7)\%$ of COMPAS observations. The full original endpoint and display-stability tables remain in the archive; the more comprehensive same-predictor checks above cover both references and all four routed predictors.

Table 16: Original top-five-absolute-then-supporting selection under mean/mode replacement: signed drop Δ, ten-mask count-matched random drop, paired excess and same-count single-field-occlusion ranking (Occ.), mean(seed SD). Negative and empty sets remain in averages. Neg. is $\Delta < -10^{-6}$; Empty and $|S|$ give selection behavior. Occ. uses the evaluated model/reference, not independent ground truth. Probability scales and selected counts differ across models.

Dataset	Model	Δ	Random	Excess	Occ.	Neg.%	Empty%	$\lvert S\rvert$
HELOC	H	.271(.015)	.035(.003)	.236(.015)	.292(.024)	.3	.0	3.97
HELOC	S	.272(.012)	.036(.002)	.236(.011)	.311(.003)	.0	.0	4.04
HELOC	EBM	.238(.021)	.031(.004)	.207(.018)	.272(.021)	.8	.0	3.73
HELOC	Additive NN	.286(.040)	.031(.005)	.255(.036)	.315(.034)	.4	.0	3.64
HELOC	CatBoost	.248(.019)	.032(.004)	.217(.015)	.277(.019)	5.6	.0	3.74
Shoppers	H	.260(.112)	.118(.034)	.142(.080)	.412(.050)	7.0	.0	4.41
Shoppers	S	.285(.085)	.117(.019)	.167(.067)	.422(.030)	2.3	.0	4.55
Shoppers	EBM	.345(.015)	.062(.001)	.283(.016)	.369(.013)	.4	.0	3.15
Shoppers	Additive NN	.340(.012)	.056(.006)	.284(.007)	.367(.019)	.1	.0	3.07
Shoppers	CatBoost	.360(.021)	.083(.004)	.277(.024)	.371(.022)	1.6	.1	3.28
COMPAS	H	.159(.036)	.032(.007)	.127(.030)	.204(.046)	.5	.7	4.28
COMPAS	S	.148(.031)	.029(.010)	.118(.022)	.201(.059)	.4	3.4	3.91
COMPAS	EBM	.193(.017)	.038(.005)	.155(.017)	.339(.040)	1.2	.0	3.70
COMPAS	Additive NN	.222(.025)	.030(.004)	.192(.023)	.237(.024)	1.0	.0	3.67
COMPAS	CatBoost	.194(.005)	.041(.002)	.153(.007)	.291(.018)	1.4	.0	3.80

Table 17: Hierarchical signed-response distribution. Cells average seed-specific quantiles or subgroup means, not pooled-sample quantiles. Correct/incorrect use test correctness. Low/high use each seed's empirical confidence tertiles with ties together. All predefined groups are retained. Sizes, thresholds, middle tertiles and seed SDs remain in CSVs. These descriptive groups are not validation-selected deployment thresholds.

Dataset	Q10	Median	Q90	Correct	Incorrect	Low conf.	High conf.
HELOC	.108	.260	.443	.278	.255	.184	.345
Shoppers	.013	.216	.599	.252	.332	.286	.173
COMPAS	.047	.154	.271	.169	.144	.110	.196

Table 18: Hierarchical minus single-level differences, paired within seed, mean(seed SD). Negative-fraction differences use the 0–1 scale. Shared splits do not isolate topology because regularization geometry also changes.

Dataset	ΔAUROC	ΔAUPRC	Δexcess	Δneg. fraction
HELOC	$-.0002(.0005)$	.0001(.0034)	$-.0002(.0174)$	.0026(.0045)
Shoppers	$-.0042(.0074)$	.0004(.0104)	$-.0250(.0435)$	.0469(.0645)
COMPAS	.0078(.0072)	.0045(.0051)	.0084(.0347)	.0013(.0045)

Reproduction and retained historical archive. `analysis/build_compact_appendix.py` regenerates these tables and hashes every input. The anonymous package retains all observation/seed CSVs and the long TeX reports under `archived/paper_reports/`, including the separate historical nine-dataset study with its original protocol. Supplied instructions/scripts reproduce the excluded raw datasets and checkpoints.

G Original-Cohort Post-hoc Impact and Display Sensitivity

This section retains the original three-dataset, three-seed cohort. Appendix M reports the expanded study. This supplementary analysis uses the same 36 checkpoints, both references, and 256 audit observations per run. It performs no additional training or model evaluations. Thresholds, display conditioning, and case selection are post hoc and were not preregistered.

Events and denominators. A material reversal requires $FD < 0$ and $|F|, |D| > 10^{-6}$. For each $\tau \in \{0, .001, .01, .05\}$, count its intersection with $|p(x) - p(x^{(j)})| > \tau$. For the display analysis, rank *all* original fields by $|A|$ or $|C|$, break ties by field index, and display five. The denominator is the number of actually changed fields in that display, without support filtering; it is identical at every threshold. Average field-event fractions within each observation, then observations within a run, the three datasets equally within seed, and finally the three seeds (mean and sample SD). Exclude only empty selections, whose counts are retained. These are conditional display rates, whereas the original all-changed-field rates include every actual field change; their difference cannot be attributed to a mechanism by subtraction. Each predictor–reference setting contains 2,304 observation entries. For C, mean/mode retains all 2,304 (11,520 changed display entries), while donor retains 2,302 (11,495 entries). For A, mean/mode omits 0–44 empty displays and donor omits 16–20, depending on the predictor. The original all-changed denominator has 37,121/30,261 field entries under mean/mode/donor, with 0/2 empty observations. Counts are coverage information, not aggregation weights; overlapping observations are not independent training replicates.

Table 19: Post-hoc impact among displayed fields: percentage with material F–D sign reversal and $|\Delta p| > \tau$. Rank all fields by $|A|$ or $|C|$, display five, then use the actually changed entries as the denominator. No support filtering. Observation means are averaged equally over datasets within seed. Entries are means (sample SD) across three seeds. Empty selections are omitted and their counts are exported. The four probability thresholds are illustrative, not validated usefulness criteria.

Predictor	Reference	Ranker	$\tau = 0$	.001	.01	.05
Hierarchical	Mean/mode	A	6.2 (1.3)	5.3 (0.8)	3.5 (0.5)	1.4 (0.4)
Hierarchical	Mean/mode	C	6.1 (1.1)	5.3 (1.0)	3.6 (0.7)	1.3 (0.5)
Hierarchical	Donor	A	8.5 (0.9)	7.5 (0.8)	4.9 (0.4)	2.1 (0.6)
Hierarchical	Donor	C	9.3 (1.2)	8.2 (1.2)	5.6 (0.5)	2.5 (0.5)
Single-level	Mean/mode	A	7.6 (0.5)	6.9 (0.5)	4.3 (0.7)	1.3 (0.3)
Single-level	Mean/mode	C	7.7 (0.9)	6.9 (0.7)	4.4 (0.6)	1.3 (0.3)
Single-level	Donor	A	9.8 (1.2)	8.4 (1.3)	5.2 (1.0)	1.9 (0.3)
Single-level	Donor	C	10.7 (1.2)	9.4 (1.2)	6.4 (0.8)	2.5 (0.3)
Affine dense	Mean/mode	A	10.8 (1.8)	8.3 (1.4)	4.4 (0.4)	0.9 (0.1)
Affine dense	Mean/mode	C	6.3 (0.4)	5.1 (0.4)	2.7 (0.2)	0.7 (0.1)
Affine dense	Donor	A	10.8 (2.2)	8.7 (1.8)	4.7 (0.3)	1.2 (0.0)
Affine dense	Donor	C	9.0 (0.8)	7.5 (0.5)	4.3 (0.3)	1.5 (0.1)
Affine top-two	Mean/mode	A	7.8 (2.5)	6.1 (2.4)	3.0 (1.4)	0.7 (0.3)
Affine top-two	Mean/mode	C	4.8 (1.0)	4.0 (0.8)	2.0 (0.6)	0.6 (0.2)
Affine top-two	Donor	A	8.7 (1.6)	6.7 (1.4)	3.3 (1.1)	1.2 (0.2)
Affine top-two	Donor	C	6.9 (1.8)	5.9 (1.3)	3.4 (0.9)	1.4 (0.1)

Selection sensitivity. The released tables also retain three distinct alternatives: changed fields ranked first by absolute score; the original absolute top-five display followed by positive predicted-

class support; and the main-text signed top-three selection at equal actual-change budget. The supporting-only view has 8–84 empty selections per predictor–reference setting. We verify all 73,728 saved supporting-only/top-three masks across rankers and observations. Single-field impacts here are not sums or predictions of joint-selection responses. Probability thresholds describe effect size and have no validated interpretation as usefulness cutoffs.

A real, transparently selected illustration. Among changed fields in the C absolute top-five display, require material C and F with the same sign, F and D with opposite signs, and $|\Delta p| > .05$. Across all 36 checkpoints and both references, this gives 1,236 eligible field events. Sort by $|\Delta p|$, then model, dataset, seed, reference, observation and field indices; select the lower median, rank 618. The unit is a field event, not a unique observation. This post-hoc high-impact illustration is neither a typical observation nor a maximum-effect example. The selected field is `Informational` in Affine top-two / online_shoppers, seed 1, mean_mode reference, audit index 928 (source row 7875). The reference uses the recorded fitting-set numeric mean/category mode. The calibrated positive-class probability changes from 0.5147 to 0.6049; the operating threshold is 0.3547. The original predicted label is 1, and the dataset label is 1. Here $A = 0.1576$, $C = 0.1576$, $F = 0.1576$, $R = -0.5301$, and $D = -0.3725$. Thus centering leaves the response direction wrong, while $F + R \simeq D$. Expert support is unchanged; route weights change (ℓ_1 distance 0.8350). The prediction remains correct in this case. Released endpoint records exclude raw field values; the observation and reference are identified by the recorded row and reference metadata.

Table 20: The selected observation's complete C absolute top-five display. Each row is a separate single-field replacement, not a joint intervention. Bold identifies the median-selected event. Full original-field scores, both display rankings, probabilities, exact saved joint selections, source hashes, and all candidate events are provided in the generated audit-impact records.

Field	A	C	F	R	D	$p(x^{(j)})$
PageValues	0.2277	0.2277	0.2277	-0.0087	0.2190	0.4608
Month	-0.0097	0.1584	0.1584	-0.1362	0.0222	0.5092
Informational	0.1576	0.1576	0.1576	-0.5301	-0.3725	0.6049
BounceRates	0.0923	0.0923	0.0923	0.0012	0.0935	0.4917
ExitRates	0.0815	0.0815	0.0815	0.0009	0.0824	0.4944

Reproduction. The supplement's `analysis/analyze_audit_impact.py` writes per-observation numerators and denominators, run/dataset/seed summaries, complete ordered case candidates, all fields and saved joint selections for the selected observation, and source hashes. When the optional earlier effect-size export is present, it also checks all 288 baseline summaries; the validation record states whether this check ran. A secondary hierarchical-only median illustration is included in the machine-readable records, without replacing the all-model selection above.

H Original-Cohort Full-Gradient Repair and Residual Effects

This section retains the original three-dataset, three-seed cohort; Appendix M reports the expanded study. This post-hoc extension covers all 36 saved checkpoints, the same 256 audit observations per checkpoint, and both original references. It compares $C_j = (x_j - x'_j)g_j$ with $S_j = (x_j - x'_j)\partial_j \ell(x)$ on numerical fields, differentiating through the complete original predictor. The fields are standardized numerical coordinates, identical in the router input and the numerical part of the expert input. Categorical fields are excluded: there is no derivative through their discrete labels. Thus these results have a different field denominator from the all-field audit. No model is trained or selected. All 258,048 numerical field–reference records are retained. Full and frozen derivatives are computed separately, with the latter checked against the original mechanism audit. Saved row identities, probabilities, centered scores, finite targets, and endpoint support are verified. All 864 predetermined float64 central-difference checks (two step sizes and endpoint-support checks) pass; the maximum error is 2.86×10^{-10}. The preceding HELOC dense-affine control is reproduced within floating-point

Table 21: Numerical-field repair $C \to S$, macro-averaged across the four models and three seeds within each dataset. MAE and material-target sign agreement first average changed numerical fields within each observation. Joint probability drop uses the same predicted-class signed ranking and numerical replacement budget $\min(3, n_{\text{changed}})$. Arrows show measured before/after values, not an assertion of universal improvement.

Dataset	Reference	Logit MAE ↓	Sign agreement (%) ↑	Joint drop ↑
HELOC	Mean/mode	$0.055 \to 0.022$	$85.904 \to 96.562$	$0.230 \to 0.250$
HELOC	Fit donor	$0.101 \to 0.060$	$85.849 \to 94.687$	$0.286 \to 0.303$
Shoppers	Mean/mode	$0.453 \to 0.266$	$83.301 \to 95.843$	$0.289 \to 0.321$
Shoppers	Fit donor	$0.468 \to 0.587$	$79.870 \to 93.268$	$0.198 \to 0.203$
COMPAS	Mean/mode	$0.033 \to 0.014$	$92.974 \to 98.419$	$0.139 \to 0.149$
COMPAS	Fit donor	$0.064 \to 0.039$	$95.835 \to 98.070$	$0.167 \to 0.174$

Table 22: Material-reversal-and-effect incidence (%) after complete-predictor differentiation. Average the field incidence within each nonempty observation display, then equally average the 36 checkpoint means. The first two outcome columns use exactly the same fixed C display. The last column uses a separately selected S display. Full counts, zero-score/zero-target cases, empty subsets, and alternative field-weighted rates are released.

Reference	δ	C, fixed C display	S, fixed C display	S, own display
Mean/mode	0	5.868	2.611	1.289
Mean/mode	0.001	4.774	2.046	1.092
Mean/mode	0.01	2.522	1.050	0.645
Mean/mode	0.05	0.664	0.332	0.247
Fit donor	0	9.222	4.127	3.188
Fit donor	0.001	7.022	3.431	2.561
Fit donor	0.01	3.971	1.985	1.503
Fit donor	0.05	1.399	0.719	0.599

tolerance, and all source hashes remain unchanged. One hierarchical HELOC checkpoint (seed 3) has no material local router derivative: its maximum full-minus-frozen derivative is 8.94×10^{-8}. We retain this case rather than assuming that every checkpoint needs the correction. Using a 10^{-8} tolerance to treat numerical changes as ties, S improves MAE in 60/72 checkpoint–reference pairs (2 ties, 10 worse), sign agreement in 69/72 (2 ties, 1 worse), and joint probability drop in 52/72 (2 ties, 18 worse). Across all 36 checkpoints, mean/mode logit MAE falls from 0.180 to 0.101, whereas donor MAE rises from 0.211 to 0.228, driven by Shoppers. Thus correcting the local derivative improves direction much more consistently than finite-step magnitude or joint selection. In particular, donor substitutions can move beyond the regime where a first-order approximation is accurate.

Fixed-display residual incidence. For the primary display comparison, rank $|C|$ over numerical fields, retain five fields with stable field-index tie breaking, and intersect with actually changed fields. Both C and S are evaluated on this identical set. A material reversal requires opposite signs and $|\text{score}|, |D| > 10^{-6}$. The numerator additionally requires $|p(x) - p(x'_j)| > \delta$. The denominator is the fixed display, including fields below the probability threshold; it is not the number of high-effect fields. A separate display based on $|S|$ is shown without treating its denominator as interchangeable. Thresholds are post-hoc effect-size descriptions, not validated decision-utility cutoffs. At $\delta = 0.05$, the fixed-display event counts fall from 306 to 153 among 46,080 mean/mode entries, and from 603 to 313 among 43,628 donor entries. As an alternative aggregation, first pooling fields within each checkpoint and then averaging checkpoints yields 0.664% to 0.332% for mean/mode and 1.424% to 0.733% for donors; fully pooled donor rates are 1.382% and 0.717%. These field-weighted rates

differ from the primary observation-weighted rates in Table 22 when display sizes differ. The 12 empty donor observation displays are omitted from their checkpoint observation means and retained in the released counts. These modest residual incidences should not be represented as common large-effect failures. Of the remaining S events, 129/153 under mean/mode and 204/313 under donor replacement occur when the nonzero expert support differs between the two endpoints. The other 24 and 109 events have identical endpoint support. These endpoint strata do not establish whether support changes in the interior of the replacement path. Smooth dense mixtures also retain finite-step residuals: the derivative correction is local, while D reruns the predictor at the finite endpoint. Neither these counts nor sign agreement establish causal or human-facing utility.

I Original-Cohort Budgeted Endpoint Checking

This section retains the original three-dataset, three-seed cohort; Appendix M reports the expanded study. This post-hoc, offline follow-up tests whether local scores prioritize finite endpoint checks. For each observation, fix the five largest absolute full-gradient scores $|S|$ among numerical fields, then omit no-op fields. A query replaces one displayed score by its exact finite response D, leaving the prediction and other scores unchanged. Both priorities use only original-point information and the declared reference: $|S - C|$ (routing derivative) or $|S|$ (score magnitude). Budgets are 1, 2, 3, and 5, capped by the changed display size; all are released. The matched random control uses the exact expected remaining error $(1 - b/n)$ times the unchecked total, so there is no Monte Carlo variability. This is a field-query budget, not a wall-clock speedup claim. The common display is selected before checking and is not reranked after correction. Error rates average nonempty displays; all 12 empty displays are counted and omitted from those means. Average query counts include the zero cost of empty displays. These rates concern numerical fields and cannot be subtracted from the earlier all-field display rates. All per-observation masks, unqueried residuals, logit MAE, near-zero-score counts, thresholds and budgets are supplied. Population significance, human usefulness and external-system prevalence are not established.

Table 23: Budgeted checks on the fixed numerical S display. Residual material direction reversals jointly with $|\Delta p| > .05$, as percentages of displayed changed fields. Three-seed mean $\pm$ SD after averaging observations and datasets. The three checking policies each query up to two fields. Exact queried responses remove error by construction. Prioritization is useful only relative to the matched controls.

Model	Ref.	Unchecked	Random	$\lvert S - C\rvert$	$\lvert S\rvert$
Hierarchical	Mean/mode	0.37 ± 0.25	0.22 ± 0.15	0.10 ± 0.03	0.20 ± 0.08
Hierarchical	Donor	0.73 ± 0.17	0.41 ± 0.10	0.31 ± 0.14	0.34 ± 0.14
Single-level	Mean/mode	0.11 ± 0.07	0.07 ± 0.04	0.03 ± 0.04	0.04 ± 0.02
Single-level	Donor	0.58 ± 0.22	0.32 ± 0.12	0.19 ± 0.07	0.29 ± 0.07
Affine dense	Mean/mode	0.06 ± 0.02	0.04 ± 0.01	0.03 ± 0.03	0.03 ± 0.03
Affine dense	Donor	0.37 ± 0.10	0.21 ± 0.06	0.05 ± 0.03	0.18 ± 0.08
Affine top-two	Mean/mode	0.44 ± 0.08	0.27 ± 0.05	0.23 ± 0.18	0.36 ± 0.14
Affine top-two	Donor	0.72 ± 0.17	0.41 ± 0.10	0.26 ± 0.13	0.36 ± 0.06

J Original-Cohort Repeated Donor and Local-Reference Sensitivity

This section retains the original three-dataset, three-seed cohort; Appendix M reports the expanded study.

This post-hoc follow-up keeps all 36 original checkpoints and their 256 audit observations. For each dataset and training seed, five independent whole-row donors are drawn from the fitting subset and shared across all four predictors. We compare original A and centered C at the identical actual-change

Table 24: Post-hoc donor sensitivity at the same actual-change budget $k = 3$. Each training-seed result first averages five donor draws. Entries are paired $C - A$ joint probability response in percentage points (mean $\pm$ sample SD over three training seeds). Donor SD is the mean within-checkpoint sample SD over five draws. Draws are shared across four predictors. Repeated references and overlapping holdouts are not independent population replicates.

Dataset	Predictor	Random donor $C - A$	Donor SD	Nearest-ten $C - A$
HELOC	Hierarchical	6.34 ± 1.60	0.50	4.62 ± 0.75
HELOC	Single-level	5.88 ± 0.89	0.54	4.73 ± 0.76
HELOC	Affine dense	7.20 ± 0.58	0.29	4.82 ± 0.22
HELOC	Affine top-2	6.46 ± 0.85	0.57	4.72 ± 0.23
Online Shoppers	Hierarchical	5.39 ± 0.85	1.03	–
Online Shoppers	Single-level	4.55 ± 0.67	1.07	–
Online Shoppers	Affine dense	9.52 ± 1.19	0.93	–
Online Shoppers	Affine top-2	8.21 ± 1.59	0.99	–
COMPAS	Hierarchical	4.54 ± 1.02	0.47	–
COMPAS	Single-level	4.21 ± 0.94	0.41	–
COMPAS	Affine dense	3.43 ± 0.20	0.31	–
COMPAS	Affine top-2	3.38 ± 0.25	0.32	–

budget $\min(3, n_{\text{changed}})$, retaining negative-ranked scores if necessary and recomputing the complete joint response. The original donor draw is excluded from the new aggregate.

For HELOC, five additional draws sample uniformly from each observation's ten nearest fitting rows in the original 23 standardized numerical coordinates; distance ties use source-row order. Special codes retain their original numerical treatment. Neighbors are selected on the whole input, including the field later replaced, so this is a distance-based sensitivity check, not a conditional or causal intervention. Partial replacements may remain atypical.

We average observations within each donor condition, donor draws within a training seed, and only then report means and sample SD over the three training seeds. No inferential claim treats donor draws as additional training seeds. The accompanying records retain every checkpoint, draw, actual budget and source-row identity.

Random fitting-row donors favor C in 180/180 condition means, and in all five draws for 36/36 checkpoints. The paired condition contrasts range from 2.72 to 11.69 percentage points. These are descriptive reference-sensitivity counts.

HELOC nearest-ten donors favor C in 60/60 condition means, and in all five draws for 12/12 checkpoints. The paired condition contrasts range from 3.48 to 5.86 percentage points. These are descriptive reference-sensitivity counts.

A secondary analysis retains single-field F–D reversals with $|F|, |D| > 10^{-6}$ and absolute calibrated probability changes exceeding $0, .01, .05$, for all changed fields and changed entries among the five largest $|C|$. These rates use fixed denominators across thresholds; empty display subsets are omitted and counted. Full results, including every unfavorable condition, are retained in the CSVs.

For dense affine mixtures under repeated random donors, macro reversal rates above a .05 probability-change threshold are $1.73 \pm 0.09\%$ across all changed fields and $1.33 \pm 0.11\%$ within the centered top-five display. On HELOC alone, that display rate is 1.20% under random donors and 0.17% under nearest-ten donors. Effect magnitude remains reference dependent; the references have different changed-field denominators.

K External SENN Reproduction and Logit-Aligned Control

Source and native target. We use the original authors' SENN repository, pinned to commit `5a7907aefa926877a57cb9c112f7942442b1714b`. The release includes the unmodified

source, its copyright notices, a separate compatible copy, and an exact diff. Compatibility edits remove unused unavailable imports and modernize scalar logging; they do not alter the model or released regularizer. Our experiment reuses the official identity conceptizer, relevance network, binary additive aggregator, and `GradPenaltyTrainer(typ=3)`. It adapts the evaluation harness and is not an exact rerun of the original paper's benchmark search.

SENN exposes concept–coefficient pairs (h_i, θ_i); `explain()` returns θ, while $\theta_i h_i$ is a derived additive ledger. With $h = (x, 1)$ and pre-sigmoid score $\ell = \theta^\top h$, the ledger includes the bias concept and sums to ℓ. Neither the raw coefficients nor ledger terms claim exact arbitrary finite responses. The primary native test compares numerical θ_j with $\partial_j \ell$; a companion probability-scale test compares $p(1-p)\theta_j$ with $\partial_j p$, not raw θ_j with a probability derivative. The bias concept is never an input field.

Two fixed training arms. Both arms use HELOC, Adult, and MAGIC, seeds 1–5, the shared fitting-only 64/16/20 splits, and the same architecture: an identity conceptizer with bias and a tanh relevance network of hidden widths $(\lfloor d/2 \rfloor, 10, 5)$. Categorical inputs use fitting-only one-hot vocabularies, including unknown codes, retaining the original field groups. They affect predictions, but the derivative and finite-response audit is restricted to numerical fields. Thus "native top-five" below means the five largest absolute coefficients *within the numerical subset*, not a display ranking all numerical and categorical concepts together.

Training uses Adam at .001, batch size 128, 40 epochs, regularization weight .01, and minimum validation BCE (earliest tied epoch). The optimizer follows the released implementation without weight decay. We replace the original UCI script's broad hyperparameter grid and validation-accuracy selection with this fixed harness; no test-based model, dataset, arm or regularization choice is made. Both arms retain the native sigmoid and threshold .5, without fitted calibration. Full-test predictive metrics are recomputed from saved predictions.

In the released arm, `GSENN.forward` detaches the identity concepts. Its `grad3` penalty differentiates the returned sigmoid probability through that graph, pads a zero bias derivative, and compares it with the coefficient vector using the batch L2/Frobenius norm. We preserve these operations, including nonzero regularization. For the separate logit-aligned arm, choose $f = \ell$ and use

$$\mathcal{L}_{\text{aligned}} = \text{BCE}(p, y) + .01 \left\| \nabla_x \ell - \theta_{\text{nonbias}} \right\|_F . \tag{17}$$

The identity chain is differentiable and $J_x h = [I; 0]$, so the constant concept contributes no coordinate derivative. This Eq. 3-style control changes the link scale, detached identity chain, and bias treatment together. It does not isolate their individual effects or establish the original authors' unique intended classification scale. The fixed execution identifier `paper_aligned_logit` is retained for provenance; the manuscript calls the arm *logit-aligned*. This is an explicit experimental control, not a compatibility-only patch or official corrected release.

Same-function audit. For each trained checkpoint, reconstruct the mathematical input dependence using the same learned coefficient network and differentiable identity concepts. Check equality with the native forward at original and replaced inputs. This changes the differentiation graph, not the function values or weights. All 2,880 predetermined float64 central-difference checks pass, with maximum error 3.19×10^{-10}. The original 256-row audit RNG and donor RNG are shared with the routed study; split hashes, fitting membership and donor identities are independently reconstructed.

For numerical displacement $\Delta x_j = x_j - x'_j$, define the derived response $C_j = \Delta x_j \theta_j(x)$, complete local response $S_j = \Delta x_j \partial_j \ell(x)$, and endpoint target $D_j = \ell(x) - \ell(x')$. Since identity concepts are affine,

$$D_j = C_j + [\theta(x) - \theta(x')]^\top h(x'). \tag{18}$$

The second term includes the bias coefficient's response. Its maximum measured reconstruction residual is 6.97×10^{-6}. SENN coefficients need not be positive or normalized, so the mixture probability-simplex bound in Appendix B does not apply to this model.

The displayed incidence divides material sign reversals with the stated probability effect by the same actually changed numerical display fields, including fields below the effect threshold. Empty displays and all denominators are released. Joint tests rank predicted-class signed C, S, D at equal actual budget $\min(3, n_{\text{changed}})$ and recompute every joint endpoint. Scores based on single-field D are a comparator, not a joint combinatorial optimum. The released arm's Adult joint drop slightly decreases

Table 25: SENN donor-reference practical-impact checks (mean ± seed SD). Both C and S use the identical top-five absolute native numerical coefficients, with actual-change denominators fixed across effect thresholds. Native sigmoid probability changes are uncalibrated. Joint rankings use the same actual-change budget $\min(3, n_{\text{changed}})$, predicted-class signed scores, and independently recomputed joint endpoints; D ranking is not a combinatorial optimum.

		Native top-five reversals (%)		Donor joint $k = 3$ probability drop		
Dataset	Training arm	C, $\lvert\Delta p\rvert > .05$	S, $\lvert\Delta p\rvert > .05$	C	S	D
HELOC	Released	0.073 ± 0.072	0.031 ± 0.043	0.1927 ± 0.0104	0.2196 ± 0.0132	0.2350 ± 0.0090
HELOC	Logit-aligned	0.000 ± 0.000	0.000 ± 0.000	0.3532 ± 0.0211	0.3532 ± 0.0212	0.3535 ± 0.0212
Adult	Released	2.743 ± 0.789	2.413 ± 1.088	0.1290 ± 0.0106	0.1283 ± 0.0087	0.1383 ± 0.0092
Adult	Logit-aligned	0.000 ± 0.000	0.000 ± 0.000	0.1273 ± 0.0183	0.1273 ± 0.0183	0.1273 ± 0.0183
MAGIC	Released	1.141 ± 0.617	0.766 ± 0.289	0.3994 ± 0.0099	0.4491 ± 0.0163	0.4615 ± 0.0124
MAGIC	Logit-aligned	0.000 ± 0.000	0.000 ± 0.000	0.5556 ± 0.0207	0.5549 ± 0.0212	0.5580 ± 0.0200

Table 26: Donor-path step-size sensitivity for SENN (mean ± seed SD). Errors are $|C - D|/|\Delta x_j|$ or $|S - D|/|\Delta x_j|$, averaged over actually changed numerical fields within each observation and then observations. Endpoint interpolation uses the fixed fractions $t = .001, .01, .1, 1$. Intermediate fractions and the mean-reference results remain in the accompanying records. Small-step and full-step results evaluate different target regimes, not a native finite-change claim. Endpoints and neural evaluations use float32. Subtractive cancellation and rounding limit the smallest steps, so these numbers are not a numerical verification of the zero-step limit.

		C: normalized MAE		S: normalized MAE	
Dataset	Training arm	$t = .001$	$t = 1$	$t = .001$	$t = 1$
HELOC	Released	0.0953 ± 0.0063	0.0952 ± 0.0064	0.0002 ± 0.0000	0.0189 ± 0.0023
HELOC	Logit-aligned	0.0059 ± 0.0012	0.0065 ± 0.0011	0.0003 ± 0.0000	0.0030 ± 0.0004
Adult	Released	0.2299 ± 0.0211	0.2482 ± 0.0199	0.0006 ± 0.0003	0.0939 ± 0.0084
Adult	Logit-aligned	0.0005 ± 0.0003	0.0006 ± 0.0007	0.0004 ± 0.0001	0.0005 ± 0.0007
MAGIC	Released	0.4317 ± 0.0257	0.4045 ± 0.0207	0.0010 ± 0.0002	0.1306 ± 0.0138
MAGIC	Logit-aligned	0.0297 ± 0.0121	0.0740 ± 0.0069	0.0012 ± 0.0002	0.0634 ± 0.0047

from C to S, as does the aligned MAGIC arm; improved individual directions need not improve joint selection. Comparisons of response magnitudes across retrained arms are not common-scale explanation-quality comparisons.

Fitting means and fitting-row donors define separate reference paths with fractions $.001, .01, .1, 1$. All fractions and both references are retained. Dividing absolute logit error by the actual nonzero displacement prevents smaller steps from appearing accurate solely because of smaller units. However, the endpoint evaluations use float32: very small representable displacements can yield a rounded zero logit difference and amplified normalized error. These are finite-precision step-size checks, not a proof of the limiting derivative error or a decomposition of all residual error into curvature. The independent float64 gradient checks serve the separate derivative-verification purpose.

The logit-aligned control has no observed $|\Delta p| > .05$ reversal events in the fixed native numerical display on these 15 models. It nevertheless retains finite magnitude error, particularly on MAGIC, and has lower mean AUROC on Adult and MAGIC. We report all checkpoints rather than selecting the more favorable implementation. This extension establishes a reproducible issue and measured

repairs in one released external model configuration; it does not establish failure of SENN as a general principle, learned-concept SENN variants, domain-shift robustness, causal validity, or human benefit.

L Dataset Coverage and Protocol Definitions

The main study contains ten datasets, five training seeds, and four routed predictors per dataset. Table 27 records coverage of the current experiments and the separately archived protocol. All ten datasets remain in the primary aggregate, including adverse predictive and diagnostic outcomes.

Table 27: Dataset coverage by experiment. Entries give training seeds per dataset, not numbers of observations or models. The routed study has four predictors per seed (200 checkpoints). The SENN experiment has two arms per seed (30 checkpoints). A dash means not run under that protocol. Historical and current results are not pooled.

Dataset	Historical study	Routed study	SENN experiment
Zhongyuan	20	5	–
Taiwan	20	5	–
HELOC	20	5	5
e-shop / Shoppers	20	5	–
Telco	20	5	–
Bank	20	5	–
Covertype (binary)	20	5	–
COMPAS	20	5	–
Adult	20	5	5
MAGIC	–	5	5

Why the protocols remain separate. The archived preprocessing fits transformations on the outer training partition before separating the inner validation partition. Some feature types also depend on integer/cardinality heuristics applied before splitting, and an unknown ordinal category shares code zero with a known category. Current routed experiments fix semantic types and fit preprocessing only on the inner fitting partition, with a separate unknown-category code. For example, the archived HELOC schema has 2 numerical and 21 categorical fields; the current schema has 23 numerical fields. The corresponding archived/current numerical–categorical counts are 5/9 versus 7/7 for Shoppers, 3/18 versus 12/9 for COMPAS, 1/13 versus 6/8 for Adult, and 5/14 versus 9/10 for Bank. Covertype now uses all 54 predictor fields after source verification and positional header repair; the historical loader omitted three real terrain fields. These changes prevent interpreting historical measurements as runs of the current protocol or attributing historical/current differences to a single change.

Retained evidence and limits. The anonymous archive includes the nine-dataset source manifest, task definitions, per-seed predictive records, and the historical reliability analysis. Historical Taiwan seeds 7, 13, and 19 produce constant probabilities and missing diagnostic components; predictive summaries retain them, while paired diagnostic comparisons use the declared finite subset. Those historical runs are separate from the current five-seed audit. The current Taiwan numerical failures and uniform all-seed replacement protocol are disclosed separately in Appendix O; the table counts selected checkpoints, not all attempts. Covertype is type 2 versus the remaining types, not seven-class classification. The initial 36-checkpoint audit, subsequent six-dataset audit, and present ten-dataset audit are separately recoverable. Appendix M documents source limitations and the newly completed conditions; Appendix N evaluates all one- and two-dataset omissions without selecting a preferred subset.

Names of the controlled cohorts. In the following tables, "Initial" means HELOC/Shoppers/COMPAS, "Cohort B" means Adult/Bank/MAGIC, and "Restored" means Zhongyuan/Taiwan/Telco/Covertype. Machine-readable cohort identifiers containing `original` or

Table 28: Source schemas for Adult, Bank, and MAGIC. Positive labels are income above $50K, deposit subscription, and simulated gamma events, respectively. Counts use the original source before splitting. All five fixed seeds pass source-identity, disjoint-split, repeatability, and held-out-mutation checks.

Dataset	Rows	Numerical	Categorical	Fit / val / test
Adult	32,561	6	8	20,838 / 5,210 / 6,513
Bank	41,188	9	10	26,360 / 6,590 / 8,238
MAGIC	19,020	10	0	12,172 / 3,044 / 3,804

Table 29: Source schemas for Zhongyuan, Taiwan, Telco, and Covertype under fitting-only preprocessing. Counts and field schemas are checked against every completed training and audit manifest. Positive classes are default (Zhongyuan and Taiwan), churn (Telco), and cover type 2 versus the other six types (Covertype).

Dataset	Rows	Numerical	Categorical	Fit / val / test
Zhongyuan	10,000	9	22	6,400 / 1,600 / 2,000
Taiwan	30,000	14	9	19,200 / 4,800 / 6,000
Telco	7,043	3	16	4,507 / 1,127 / 1,409
Covertype	581,012	10	44	371,847 / 92,962 / 116,203

`new_datasets` retain their execution-time names for provenance. They identify controlled subsets rather than dataset novelty or the historical nine-dataset boundary. Their overlapping results are not independent replications.

M Dataset Schemas, Configurations, and Detailed Results

The controlled study comprises Zhongyuan, Taiwan, HELOC, Shoppers, Telco, Bank, Covertype, COMPAS, Adult, and MAGIC: ten datasets, five seeds, and four routed predictors, giving 200 checkpoints. All datasets and declared conditions are retained. Appendix L maps execution cohorts to their fixed protocols; those identifiers are preserved in the machine-readable records.

The initial 36-checkpoint analyses and the subsequent 120-checkpoint six-dataset analyses remain separately recoverable. The four restored datasets contribute 80 routed checkpoints and 20 CatBoost fits. Predictive context comprises 44 CatBoost fits: three seeds on each of HELOC/Shoppers/COMPAS and five on each remaining dataset. All routed models use five seeds. Overlapping holdouts, repeated fields, and overlapping cohorts are not independent population replicates. All learned preprocessing uses fitting rows.

Zhongyuan retains the archived 10,000-row table and its previously audited field set, excluding identifiers and the three early-repayment fields. Its public provenance and feature-availability metadata remain incomplete; this benchmark is not evidence of prospective lending performance. Taiwan retains nonpositive repayment-status codes categorically (Yeh, 2009). Its complete table matches the official UCI workbook; 27 differing bytes encode only a saved-by marker and modification time. The released source-equivalence report distinguishes the public workbook hash from the archived hash without disclosing identifying workbook metadata. Telco uses the IBM sample table (IBM, 2019): all 7,043 rows and 21 source columns match the pinned IBM file, with only LF versus CRLF line endings differing. Its 11 blank TotalCharges entries are retained and imputed with the fitting-subset median. This sample benchmark does not establish operational customer validity. These three datasets retain 31, 23, and 19 predictor fields, respectively.

Covertype uses all 581,012 rows (Blackard, 1998) and the binary cover-type-2-versus-rest task, not the original seven-class classification task. The archived CSV header mislabels the first 14 positions:

the three leading unnamed columns contain real measurements, not identifiers. Before training, the frozen loader assigns the canonical positional schema, restoring Elevation, Aspect, and Slope and retaining 54 predictors (10 numerical and 44 categorical indicators). Source bytes and row order remain unchanged. These corrected runs are not a like-for-like comparison with historical 51-field results. Wilderness and soil indicators remain separate fields; replacing one indicator can violate a one-hot group, so those probes describe fixed-model responses rather than valid ecological configurations. Numerical-gradient results concern the ten quantitative fields. Random row splits do not establish geographic generalization.

Adult uses the 32,561-row training-size source, not the combined 48,842-row collection (Becker & Kohavi, 1996). Its row multiset matches the official file; a released integer row permutation and recovery script reconstruct our existing CSV byte-for-byte from the official source. Bank uses `bank-additional-full`, excluding call duration under the declared pre-call task (Moro et al., 2014); random splits do not test prospective temporal generalization. Its `pdays=999` sentinel remains numerical. MAGIC contains CORSIKA Monte Carlo events (Bock, 2004), not observed telescope measurements; its supplied hadron prevalence differs from deployment. Bank and MAGIC match official source hashes exactly. These datasets broaden benchmark coverage without establishing transportability to a new operational population.

The Adult/Bank/MAGIC routed runs reuse the frozen model, objective, optimizer, and audit implementations, including the existing dataset/default configuration dispatch; this does not mean identical hyperparameter values across datasets. Runtime diffs change only the loader import and accepted dataset names. Adult/Bank use the original small-benchmark configuration (routing width 72, encoder width 96, expert hidden widths 96/128, dropout .08, learning rate 4×10^{-4} focal exponent .5, sampler exponent .28, and TEC/sparsity/stability/uncertainty/ranking/negative coefficients .008/.002/.004/.004/.75/.03). MAGIC enters the existing default branch: routing width 64, encoder width 96, expert hidden widths 64/96, dropout .06, learning rate 6×10^{-4}, focal exponent 1, sampler exponent .18, and TEC/sparsity/stability/uncertainty/ranking/negative coefficients .012/.003/.006/.006/.45/.09. Other objective coefficients remain .01/.001/.01 for balance/entropy/depth. These three datasets have $G = M = 3$ before flattening and batch size 256, because the runner explicitly supplies that batch size. These rules precede extension outcomes; no configuration was tuned after observing test results.

The restored hierarchical runs retain the existing dataset-specific dispatch. Taiwan and Telco use the small-benchmark settings above. Covertype uses $G = M = 4$, routing width 96 and encoder width 128, with the same .08 dropout, 96/128 nonlinear-expert widths, learning rate 4×10^{-4} and small-benchmark objective coefficients. Zhongyuan uses $G = M = 4$, routing width 80, encoder width 96, nonlinear-expert widths 96/112, dropout .06, learning rate 4×10^{-4}, focal exponent 1 and sampler exponent .12; its TEC/sparsity/stability/uncertainty/ranking/negative coefficients are .012/.003/.006/.006/.55/.14. All use batch size 256. Flattening preserves the dataset-specific expert-bank size and active budget. These fixed recipes were not tuned against the restored test outcomes.

The same 256 test observations, two references, $k = 1, 3, 5$ actual-change budgets, and ten matched random controls are used. Full-gradient records exclude categorical labels, retain every numerical field, and verify saved endpoints and finite differences. Frozen source hashes, predictions, split identities, training logs and compatibility diffs are recorded. The added-seed, added-to-audit, and restored-dataset cohorts are reported separately before the complete ten-dataset aggregate. The restored datasets use the frozen fitting-only loader and existing model configuration dispatch; source/task definitions and reconstruction details accompany the released runtime. No historical raw result is substituted for a missing run under the controlled protocol.

Displayed-field event counts. Across the ten-dataset cohort, complete differentiation reduces material direction reversals above a .05 probability change from 6,135 to 3,662 among 245,760 mean/mode display entries, and from 9,177 to 6,276 among 233,768 donor entries. Of the remaining events, 1,941/3,662 and 3,432/6,276 have different endpoint supports; identical endpoint support does not exclude intermediate switches. These pooled counts and the observation-weighted rates have different denominators when display sizes vary. All 32 empty donor displays remain in the corresponding observation summaries.

Table 30: Average precision (AUPRC), mean ± seed SD. Every routed entry uses five seeds; CatBoost uses three on HELOC/Shoppers/COMPAS and five on Adult/Bank/MAGIC/Zhongyuan/Taiwan/Telco/Covertype. Training objectives and tuning budgets differ. MAGIC values concern the supplied simulated class prevalence.

Dataset	Hierarchical	Single-level	Affine dense	Affine top-2	CatBoost
HELOC	0.778 ± 0.014	0.780 ± 0.013	0.776 ± 0.010	0.776 ± 0.011	0.791 ± 0.002
Shoppers	0.706 ± 0.027	0.702 ± 0.021	0.706 ± 0.037	0.709 ± 0.022	0.739 ± 0.023
COMPAS	0.696 ± 0.010	0.698 ± 0.018	0.705 ± 0.018	0.696 ± 0.016	0.732 ± 0.005
Adult	0.777 ± 0.008	0.776 ± 0.009	0.782 ± 0.007	0.781 ± 0.009	0.826 ± 0.006
Bank	0.458 ± 0.013	0.458 ± 0.015	0.455 ± 0.020	0.461 ± 0.013	0.472 ± 0.010
MAGIC	0.955 ± 0.002	0.954 ± 0.004	0.951 ± 0.002	0.949 ± 0.003	0.958 ± 0.002
Zhongyuan	0.261 ± 0.013	0.269 ± 0.027	0.265 ± 0.018	0.256 ± 0.018	0.303 ± 0.021
Taiwan	0.535 ± 0.011	0.537 ± 0.013	0.543 ± 0.016	0.540 ± 0.012	0.551 ± 0.008
Telco	0.650 ± 0.013	0.649 ± 0.012	0.653 ± 0.003	0.646 ± 0.011	0.664 ± 0.012
Covertype	0.925 ± 0.009	0.929 ± 0.018	0.934 ± 0.003	0.928 ± 0.002	0.905 ± 0.001

Table 31: Numerical-field $C \to S$ repair by dataset, averaging the four model means within each seed and then five seeds. Reversals use the fixed C display and $|\Delta p| > 0.05$. All means, SDs, effect thresholds, support strata, near-zero cases, and empty displays are available in the numerical release.

Dataset	Reference	Logit MAE	Sign (%)	Reversal (%)
HELOC	Mean/mode	$0.062 \to 0.024$	$85.0 \to 96.6$	$0.809 \to 0.234$
HELOC	Fit donor	$0.108 \to 0.062$	$84.7 \to 94.7$	$1.739 \to 0.853$
Shoppers	Mean/mode	$0.429 \to 0.251$	$82.9 \to 95.7$	$1.367 \to 0.566$
Shoppers	Fit donor	$0.462 \to 0.598$	$79.6 \to 92.9$	$2.591 \to 1.136$
COMPAS	Mean/mode	$0.034 \to 0.014$	$93.4 \to 98.6$	$0.164 \to 0.133$
COMPAS	Fit donor	$0.066 \to 0.040$	$95.8 \to 98.2$	$0.367 \to 0.285$
Adult	Mean/mode	$0.186 \to 0.094$	$87.2 \to 95.9$	$1.629 \to 1.250$
Adult	Fit donor	$0.405 \to 0.284$	$84.8 \to 92.7$	$3.824 \to 2.729$
Bank	Mean/mode	$0.092 \to 0.046$	$85.0 \to 95.5$	$0.926 \to 0.363$
Bank	Fit donor	$0.156 \to 0.110$	$82.4 \to 92.9$	$1.715 \to 0.936$
MAGIC	Mean/mode	$0.557 \to 0.444$	$79.6 \to 89.1$	$4.254 \to 4.188$
MAGIC	Fit donor	$0.731 \to 0.716$	$77.3 \to 86.1$	$7.531 \to 7.082$
Zhongyuan	Mean/mode	$0.053 \to 0.028$	$86.4 \to 95.3$	$1.059 \to 0.387$
Zhongyuan	Fit donor	$0.074 \to 0.049$	$85.8 \to 94.1$	$1.527 \to 0.746$
Taiwan	Mean/mode	$0.055 \to 0.022$	$82.5 \to 96.6$	$0.910 \to 0.301$
Taiwan	Fit donor	$0.078 \to 0.050$	$81.8 \to 95.7$	$1.912 \to 0.988$
Telco	Mean/mode	$0.202 \to 0.075$	$90.4 \to 97.6$	$1.777 \to 0.326$
Telco	Fit donor	$0.244 \to 0.125$	$90.3 \to 97.0$	$2.266 \to 0.566$
Covertype	Mean/mode	$0.656 \to 0.716$	$72.8 \to 87.7$	$11.781 \to 6.687$
Covertype	Fit donor	$0.855 \to 1.095$	$71.6 \to 84.8$	$14.609 \to 10.281$

Numerical gradient validation. Of 4,680 predetermined checks, 4,676 pass and 4 are flagged as piecewise boundaries. The maximum passed absolute discrepancy is 2.80×10^{-8}. The audit retains 1,054,720 numerical field–reference records. These software checks verify the implemented derivative, not finite-step explanation quality.

M.1 Grouped donor sensitivity on Covertype

This post hoc sensitivity retains all 20 fixed Covertype checkpoints (four models, five seeds) and the same 256 observations and fitting-row donors per checkpoint. The 54 fields form 12 replacement

Table 32: Paired centering gains $C - A$ in joint signed probability drop, in percentage points, at equal actual-change budget $k = 3$: mean $\pm$ SD over five seeds. Negative and zero outcomes are retained. Part 1 of 2. The plot shows every individual difference.

Dataset	Predictor	Mean/mode	Fit donor
HELOC	Hierarchical	0.00 ± 0.00	5.76 ± 1.11
HELOC	Single-level	0.00 ± 0.00	6.20 ± 0.84
HELOC	Affine dense	0.00 ± 0.00	6.58 ± 0.67
HELOC	Affine top-2	0.00 ± 0.00	5.97 ± 0.88
Shoppers	Hierarchical	6.03 ± 6.90	7.33 ± 1.30
Shoppers	Single-level	5.70 ± 4.85	7.21 ± 1.34
Shoppers	Affine dense	12.83 ± 6.53	9.75 ± 4.38
Shoppers	Affine top-2	9.92 ± 1.99	10.20 ± 1.65
COMPAS	Hierarchical	0.34 ± 0.38	3.92 ± 0.62
COMPAS	Single-level	0.12 ± 0.57	3.64 ± 0.85
COMPAS	Affine dense	0.17 ± 0.14	3.64 ± 0.80
COMPAS	Affine top-2	0.21 ± 0.12	3.48 ± 0.60
Adult	Hierarchical	1.51 ± 1.77	5.48 ± 0.75
Adult	Single-level	5.39 ± 5.50	6.73 ± 2.83
Adult	Affine dense	1.72 ± 1.44	5.81 ± 1.33
Adult	Affine top-2	1.72 ± 0.85	5.82 ± 1.86
Bank	Hierarchical	0.80 ± 1.60	3.12 ± 2.06
Bank	Single-level	0.66 ± 0.93	4.82 ± 2.66
Bank	Affine dense	0.76 ± 0.34	2.72 ± 0.90
Bank	Affine top-2	0.92 ± 0.48	2.91 ± 0.49

units: ten numerical singletons, one four-indicator wilderness block, and one forty-indicator soil block. Whole-block replacements preserve one-hot validity, without establishing ecological or conditional plausibility. Group scores A_G and C_G sum their field scores; F_G and D_G are recomputed from the complete single-group endpoint. D_G is a single-group response benchmark, not an attribution ground truth or an optimizer of joint selections. Across the 20 checkpoints, the $k = 3$ paired $C - A$ gain is 5.527 ± 2.070 pp; 19/0/1 checkpoints have positive/zero/negative gains (zero tolerance 10^{-8} in probability units). All 20 signed gains, per-model summaries, and $k = 1, 3, 5$ results are released. Of 5120 observation–checkpoint pairs, 5120 have at least one material changed-group target and 0 have none; the latter have undefined sign agreement. MAE is defined for 5120 pairs; 0 have no changed group. Material targets satisfy $|D_G| > 10^{-6}$; a correct sign also requires score magnitude above 10^{-6}. These counts describe the metric denominators, not independent statistical samples. Undefined checkpoints are never silently dropped from a model/seed macro. For every method, selection uses $\min(k, \#\text{changed groups})$ groups and recomputes the joint endpoint. A soil group may replace forty fields, so these budgets are reported separately from the primary single-field audit. No confidence intervals or claim of joint optimality are attached to this descriptive sensitivity.

N Sensitivity to Dataset Composition

The ten-dataset cohort remains primary. We enumerate all ten leave-one-dataset-out and all 45 leave-two-datasets-out subsets, without choosing a favorable omission. Within each seed, the four checkpoint means and retained datasets receive equal weight; the table averages all five seeds. Positive values denote larger joint signed probability drop, better material-sign agreement, or lower logit MAE, respectively. Centering compares all-field $C - A$; gradient effects compare numerical $S - C$ (reversed for MAE). The ranges describe overlapping dataset subsets, not confidence intervals or independent replication. All 56 subsets, five seed values and sample SDs are released.

Table 33: Paired centering gains $C - A$ in joint signed probability drop, in percentage points, at equal actual-change budget $k = 3$: mean $\pm$ SD over five seeds. Negative and zero outcomes are retained. Part 2 of 2. The plot shows every individual difference.

Dataset	Predictor	Mean/mode	Fit donor
MAGIC	Hierarchical	0.00 ± 0.00	8.54 ± 1.25
MAGIC	Single-level	0.00 ± 0.00	9.67 ± 2.07
MAGIC	Affine dense	0.00 ± 0.00	5.94 ± 1.91
MAGIC	Affine top-2	0.00 ± 0.00	6.56 ± 1.70
Zhongyuan	Hierarchical	1.67 ± 1.51	3.44 ± 1.73
Zhongyuan	Single-level	2.09 ± 2.02	4.61 ± 0.67
Zhongyuan	Affine dense	0.97 ± 0.45	2.40 ± 0.25
Zhongyuan	Affine top-2	0.74 ± 0.27	2.25 ± 0.32
Taiwan	Hierarchical	3.79 ± 2.82	4.40 ± 2.10
Taiwan	Single-level	3.26 ± 1.65	5.28 ± 2.96
Taiwan	Affine dense	3.70 ± 0.72	4.06 ± 0.41
Taiwan	Affine top-2	2.87 ± 1.09	3.81 ± 1.13
Telco	Hierarchical	0.55 ± 1.54	3.83 ± 0.78
Telco	Single-level	0.75 ± 2.14	4.82 ± 2.13
Telco	Affine dense	1.82 ± 0.67	5.22 ± 0.69
Telco	Affine top-2	2.01 ± 1.38	4.99 ± 0.68
Covertype	Hierarchical	0.18 ± 2.24	5.61 ± 1.57
Covertype	Single-level	1.09 ± 1.29	7.33 ± 2.82
Covertype	Affine dense	0.12 ± 0.38	5.69 ± 2.13
Covertype	Affine top-2	0.45 ± 0.23	4.47 ± 1.79

Table 34: Improving / numerically tied / worsening checkpoint means, with paired tolerance 10^{-8}. Initial = HELOC/Shoppers/COMPAS. Cohort B = Adult/Bank/MAGIC. Restored = Zhongyuan/Taiwan/Telco/Covertype. Cohorts overlap and are not independent observations. $C - A$ compares all-field joint selections. $S - C$ uses numerical fields only.

Cohort	Reference	Joint $C - A$	Sign $S - C$	MAE $S - C$
Initial, seeds 1–3	Mean/mode	21 / 12 / 3	35 / 1 / 0	35 / 1 / 0
Initial, seeds 1–3	Fit donor	36 / 0 / 0	34 / 1 / 1	25 / 1 / 10
Initial, seeds 4–5	Mean/mode	15 / 8 / 1	24 / 0 / 0	24 / 0 / 0
Initial, seeds 4–5	Fit donor	24 / 0 / 0	24 / 0 / 0	16 / 0 / 8
Initial, seeds 1–5	Mean/mode	36 / 20 / 4	59 / 1 / 0	59 / 1 / 0
Initial, seeds 1–5	Fit donor	60 / 0 / 0	58 / 1 / 1	41 / 1 / 18
Cohort B, seeds 1–5	Mean/mode	36 / 20 / 4	58 / 2 / 0	58 / 2 / 0
Cohort B, seeds 1–5	Fit donor	59 / 0 / 1	58 / 2 / 0	53 / 2 / 5
Previous six, seeds 1–5	Mean/mode	72 / 40 / 8	117 / 3 / 0	117 / 3 / 0
Previous six, seeds 1–5	Fit donor	119 / 0 / 1	116 / 3 / 1	94 / 3 / 23
Restored four, seeds 1–5	Mean/mode	69 / 0 / 11	72 / 7 / 1	59 / 6 / 15
Restored four, seeds 1–5	Fit donor	80 / 0 / 0	72 / 7 / 1	55 / 6 / 19
All ten, seeds 1–5	Mean/mode	141 / 40 / 19	189 / 10 / 1	176 / 9 / 15
All ten, seeds 1–5	Fit donor	199 / 0 / 1	188 / 10 / 2	149 / 9 / 42

O NUMERICAL FAILURE AND UNIFORM TAIWAN RERUN

The restoration audit detected nonfinite saved parameters and nonfinite raw outputs in the original Taiwan seed-4 hierarchical and single-level models. The prediction wrapper's fallback to probability .5 had concealed this failure in exported predictive scores; the mechanism audit rejected the checkpoints.

Table 35: Expanded budgeted endpoint experiment on all 200 checkpoints. Residual important reversals on the unchanged numerical S top-five display, in percent: five-seed mean $\pm$ SD after observation and dataset/model averaging. Checking policies query up to two changed fields, keep the display fixed, and replace only queried scores by their exact finite responses. The matched random rate is its exact expectation. All budgets 1/2/3/5 and per-checkpoint results are released. The original protocol is detailed in Appendix I.

Reference	Unchecked	Random	$\lvert S - C \rvert$	$\lvert S \rvert$
Mean/mode	1.302 ± 0.166	0.772 ± 0.102	0.472 ± 0.063	0.684 ± 0.061
Fit donor	2.333 ± 0.211	1.328 ± 0.122	0.844 ± 0.087	1.108 ± 0.076

Table 36: Grouped donor sensitivity. Means $\pm$ sample SD over five seeds, after equally averaging the four models within each seed. Quality uses changed groups and averages defined observation metrics within each checkpoint. The $k = 3$ column uses a budget of changed groups.

Score	MAE to D_G	Material sign (%)	$k = 3$ drop (pp)
A	1.115 ± 0.038	60.573 ± 0.428	32.463 ± 2.499
C	0.988 ± 0.057	71.417 ± 2.479	37.990 ± 3.856
F	0.935 ± 0.093	73.018 ± 3.513	38.704 ± 4.399
D	0.000 ± 0.000	100.000 ± 0.000	50.523 ± 1.592

Table 37: Full-cohort paired means and ranges over every one- and two-dataset omission. Probability/sign changes use percentage points (pp); MAE reductions use logit units. No dataset is removed from the primary analysis.

Reference	Paired effect	Full ten	Leave one: range	Leave two: range
Mean/mode	Centering joint (pp)	1.871	[1.121, 2.079]	[0.835, 2.339]
Mean/mode	Gradient sign (pp)	10.343	[9.834, 10.907]	[9.299, 11.367]
Mean/mode	Gradient MAE reduction	0.061	[0.048, 0.075]	[0.038, 0.082]
Mean/mode	Gradient joint (pp)	2.103	[1.538, 2.336]	[1.310, 2.515]
Fit donor	Centering joint (pp)	5.350	[4.987, 5.592]	[4.650, 5.867]
Fit donor	Gradient sign (pp)	9.496	[9.015, 10.287]	[8.484, 10.735]
Fit donor	Gradient MAE reduction	0.005	[-0.008, 0.032]	[-0.024, 0.053]
Fit donor	Gradient joint (pp)	1.561	[1.037, 1.734]	[0.868, 1.917]

These outputs are invalid model evaluations, not evidence of ordinary low predictive performance. Original attempt records are retained separately from the selected analysis records.

Numerically equivalent objective. The affected configurations use focal exponent $\gamma = .5$. For hard binary labels, the original factor $(1 - p_t)^\gamma$ can have a nonfinite derivative when the floating-point sigmoid rounds to the correct endpoint, despite a finite forward loss. Writing $z = (2y - 1)\ell$, we evaluate the same factor as

$$(1 - p_t)^\gamma = \exp\{-\gamma \operatorname{softplus}(z)\}, \qquad y \in \{0, 1\}.$$

The weighted BCE multiplier, auxiliary losses, data, splits, seeds, optimizer, learning rate, epoch limit, and stopping rule are unchanged. The isolated repair additionally rejects nonfinite loss, gradients, optimizer state, parameters, validation outputs, and calibrated predictions instead of silently continuing or substituting predictions. All ten repaired conditions pass these checks and both audits. Scalar checks reproduce the unstable derivative in the original expression and verify the stable

expression against its analytic derivative. They establish a concrete numerical defect; the archived training logs do not identify the first corrupting batch, so we do not claim a uniquely established causal trace for the historical failure.

Selection and accounting. Before running repaired models, we fixed a uniform replacement rule: rerun both Taiwan model families for all five seeds, including seeds whose original attempts passed, and use all ten replacements. Every other restored condition retains its original attempt; no outcome-dependent fallback is allowed. Thus the restored main cohort selects 80 routed models and 20 CatBoost controls from 110 training attempts, including the ten additional repair attempts. Combined with the preceding six-dataset cohort, the main routed analysis contains 200 selected checkpoints from 210 attempts. Failed and superseded attempts are not counted as additional effective seeds or silently overwritten. This numerical follow-up is post hoc, and is distinct from changing the dataset inclusion rule. Per-condition origin manifests bind selected records to their source code, split identities, saved-model hashes, and validation receipts.