AN OPERATOR-THEORETIC CHARACTERISATION OF LABEL NOISE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Label noise detectors inspect neighbours, model predictions, or training history, yet ask the same question: does a recorded label agree with evidence from other examples? Designing a detector means choosing what evidence to use, how examples interact, and how their agreement is judged. Exploring these choices through repeated experiments is expensive. We give a unified operator-theoretic characterisation that organises them into a common search space for detector design. The theory identifies the information needed to characterise a detector's response to a proposed change and bounds the error when its effect is approximated. This enables a constructive strategy: predict how alternatives will behave before fully evaluating them, and run exact experiments where uncertainty could change the choice. In experiments varying training and scoring while keeping features fixed, this strategy matches the selection accuracy of evaluating every alternative across all 18 cases, with up to $6.83\times$ speed-up over an optimised implementation. The result connects unification to algorithm design for label noise detection through a shared language for proposing changes and a principled way to make the search cheaper. Experimentation code: `anonymous.4open.science/r/label-noise-operator-164B`.

1 INTRODUCTION

Choosing a label noise detector also means choosing how to train it, which examples should influence it, and how to interpret its predictions. Changing any of these choices can change which labels appear suspicious. We develop an operator-theoretic characterisation for anticipating these effects and selecting configurations without training every alternative.

Nearest-neighbour methods inspect label agreement (Zhu et al., 2022), and confident learning uses predictive probabilities (Northcutt et al., 2021). Dynamics-based methods include area under the margin (AUM; Pleiss et al., 2020) and example forgetting (Toneva et al., 2019). DynaCor learns representations of training dynamics using label corruption (Kim et al., 2024), while LEMoN detects label errors using multimodal neighbours (Zhang et al., 2025). Each asks whether a recorded label agrees with evidence supplied, explicitly or through training, by other examples. Each example contributes this evidence as a *source*; a *propagation operator* determines which examples interact; a *query* maps the aggregated evidence to a score. These three roles give a common language for comparing detector mechanisms and designing changes to them (Section 2).

Figure 1 illustrates the idea for a bird image labelled *dog*. The weighted evidence from its neighbours produces a conflicting vote in a neighbour detector. A trained classifier may assign low probability to *dog*, while a dynamics detector may observe repeatedly negative margins. These choices define a structured **algorithm space of label noise detectors**: designs differ in their evidence, interactions and readouts. Our characterisation identifies the information needed to predict changes within this space, guiding which alternatives to evaluate.

For instance, this language predicts a surprising consequence of changing confidence. Two gradient comparisons respond differently depending on whether they normalise each example's contribution before combining evidence or normalise only after aggregation. At a fixed binary model state, the first is invariant to temperature; the second can reverse its interpretation even though every predicted class stays unchanged. The analysis identifies the reversal threshold and an orientation repair. Tests on

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

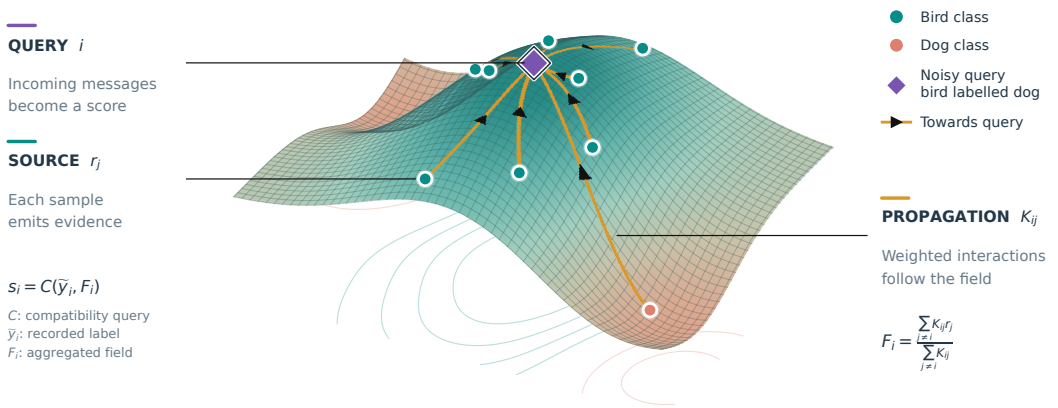


Figure 1: **Sources, propagation and queries in label-noise detection.** Sources contribute label evidence; gold paths with black arrows show weighted propagation towards a bird mislabelled as dog (purple diamond). The query aggregates incoming messages, excluding itself, and compares their support with its recorded label. This schematic Gaussian field shows bird-minus-dog evidence; the query normalises the messages before scoring conflict. The peak placement is illustrative. The framework also permits vector sources and more general interactions.

STL-10 include a separate evaluation using the gradient-comparison signal of adaptive gradient-based outlier removal (AGRA; Sedova et al., 2023; Section 6).

Predicting an intervention's effect requires an appropriate representation. Suppose we record label totals for two groups of examples and then increase the weight of their intersection. The separate group totals do not, in general, determine the label composition of that intersection. Two sources that previously produced identical training responses can consequently respond in opposite directions. Our characterisation identifies the missing information and quantifies the largest remaining ambiguity. It also specifies when enlarging the allowed interventions requires additional retained information (Section 3).

The retained information supports reusable summaries of the training evidence that predict candidate trajectories and detector scores, with approximation bounds quantifying the uncertainty in those predictions (Section 4). The forecasts support a simple search strategy: train the most promising candidates and compare their measured scores. In a prospective fixed-feature study, checking four forecast-ranked trajectories finds an optimum in all 24 search families, with median speed-ups of $5.49\times$ over exhaustive evaluation and $1.98\times$ over successive halving (Section 5).

Error bounds additionally certify selection within a prescribed tolerance. A separate 18-case study matches exhaustive selection with speed-ups up to $6.83\times$. With the forecasts unchanged, tighter error bounds reduce the additional training runs needed to certify the selected configuration from 23 to one.

How to use the algorithm space. The procedure begins by specifying the sources, interactions, query and proposed changes. A representation that retains the distinctions relevant to those changes supports prediction of their effects. Remaining uncertainty then determines which comparisons need refinement, connecting representation design to the computation required for selection.

We establish three linked contributions:

1. **Unified characterisation:** a common language for label-noise detectors that identifies what information determines their responses to proposed changes.
2. **Mechanism prediction:** a framework for analysing how detector behaviour changes, such as predicting when increased confidence reverses a detector's interpretation and how to repair it.
3. **Economical algorithm search:** predictions that guide the choice of detector–training configurations with fewer full training runs, together with bounds that can certify the choice.

Detector mechanism	Evidence and interaction	Corresponding design choice
Neighbour consistency	Recorded labels on a feature graph	Sources and neighbourhood scale
Confidence / confident learning	Fitted predictions; threshold/count stages	Predictive evidence and calibration
AUM / forgetting	Queries of an evolving margin field	Training and temporal readout
Influence / TracIn (Koh & Liang, 2017; Pruthi et al., 2020)	Curvature or checkpoint gradient responses	Reference, direction and self-effects
Co-teaching (Han et al., 2018)	Selected-source updates at fixed masks	Selection and training feedback
Fixed LEMoN	Image/text graph responses	Modalities and interactions

Table 1: **Detector mechanisms and corresponding intervention choices.** Each adaptive method is resolved into its declared evidence stages. Exact identities and state requirements appear in Section Q.

2 A Common Language for Detector Evidence

A bird recorded as a dog. Consider a bird image with the recorded label *dog*, in a representation that distinguishes birds from dogs. A neighbour detector asks whether nearby annotations support that label; a trained-model detector asks whether fitting the data makes that annotation plausible. Both combine evidence, but the messages, interactions and readout differ. Write $\widetilde{y}_i$ for the recorded label and D_i for its corruption indicator. At a declared fitted state, example j supplies a source r_j, an operator K_{ij} maps its contribution, and a query Q_i evaluates the aggregated field:

$$F_i = Z_i^{-1} \sum_{j \neq i} K_{ij} r_j, \qquad s_i = Q_i(r_i, F_i, \widetilde{y}_i). \tag{1}$$

Here Z_i specifies normalisation; local self-effects can be retained separately. Sources may contain labels, prediction residuals or histories. For neighbour voting, $r_j = e_{\widetilde{y}_j}$ and K_{ij} selects neighbours. The query compares the largest alternative class frequency with the recorded class frequency. Eight bird annotations and two dog annotations give our dog-labelled bird a conflict score of $80\% - 20\% = 60\%$.

Training interactions and detector readouts. For neural logits $f_\theta(x_i)$, let J_i be their parameter Jacobian and p_i their class probabilities. Mean cross-entropy gradient flow gives

$$\dot{f}_\theta(x_i) = -\frac{1}{n} \sum_j \underbrace{J_i J_j^\top}_{\text{propagation}} \underbrace{(p_j - e_{\widetilde{y}_j})}_{\text{source}}. \tag{2}$$

The residual of example j contributes a parameter update through $J_j^\top$, whose effect on example i is mapped by J_i. This exact finite-width identity describes the interactions underlying training-based evidence. AUM (Pleiss et al., 2020) integrates a margin trajectory, while example forgetting (Toneva et al., 2019) counts its downward decision crossings; these different readouts of the same trajectory can therefore produce different rankings. Table 1 places these mechanisms in one algorithm space.

Information retained by the source. When a binary model assigns equal probability to both labels, the residual norm is identical for either label and cannot distinguish their directions. More generally, let R_z collect the proposed label sources and $A_{u,z}$ the contributions required by query u. An operator satisfying $A_{u,z} = K_{u,z} R_z$ exists for every query precisely when

$$\ker R_z \subseteq \bigcap_u \ker A_{u,z}. \tag{3}$$

A label contrast not distinguished by the source must be irrelevant to every query. Additivity identifies the single-example responses; this factorisation characterises their source sufficiency (Section J.1). The next step asks which distinctions become relevant when we change the detector itself.

Component-specific failures and interventions. The bird example illustrates three distinct interventions. A norm-only source can be enriched with label direction; a consistently mislabelled

neighbourhood may require broader propagation to reach informative evidence; and an incorrectly oriented query requires a change to its readout. The source–operator–query account locates all three failures and specifies which components can be held fixed when testing the repair. The posterior error decomposition in Equation (21) formalises this stagewise accounting. For configuration search, we extend that principle from explaining a realised failure to predicting the responses of alternatives before fully evaluating them.

3 Information Requirements for Intervention Prediction

Extending the intervention family. Suppose we know the label totals associated with each of two feature groups. Those summaries can predict the effects of upweighting either group, but the label composition of their intersection can vary while both summaries remain unchanged. Upweighting that intersection therefore requires information that the original representation may not retain. We characterise this dependence on the intervention family explicitly.

An exact training source. Use fixed features $x = (z, c_0)$ on a compact support, a known nonzero intercept c_0, $C \geq 2$ classes and full-batch softmax training. Let μ be the feature measure and ν the joint feature–label measure, with a common objective normalisation. Known bounded weight functions span a linear space $\mathcal{H}$ containing the constant function. The label-response operator is

$$\mathcal{B}(h) = \int h(x) x e_y^\top \, d\nu(x, y).$$

This operator gives the label-weighted feature total entering training under weight h. The gradient at parameter W is

$$G_h(W) = \int h(x) x \, \operatorname{softmax}(W^\top x)^\top d\mu - \mathcal{B}(h). \tag{4}$$

The first term depends on feature geometry and the current model; the second retains the recorded-label contribution exactly.

Theorem 1 (Intervention-aware training characterisation)**.** *Under this setting, two sources produce the same gradient fields for every initial W and every admissible nonnegative weight spanning $\mathcal{H}$ if and only if their feature measures μ and operators $\mathcal{B}|_{\mathcal{H}}$ agree. These objects determine every finite training trajectory from a common initial state under the declared weights, learning rates and known ridge strengths, and every detector readout with its evaluation query fixed.*

The characterisation specifies both sufficient information and the distinctions recoverable from the complete training response. The proof identifies label totals through the intercept and the feature measure through affine sigmoid queries (Section A). For the two sign controls h_1, h_2 in the group example, $\mathcal{H} = \operatorname{span}\{1, h_1, h_2\}$ needs three label matrices. The feature measure already determines the known feature groups. Adding their interaction can require the additional matrix $\mathcal{B}(h_1 h_2)$.

Intervention compatibility under compression. Each entry of $\mathcal{B}(h)$ integrates a class indicator against a scalar test $h(x)x_k$. With fixed μ, let $V \subset L^1(\mu)$ be the closed span of the retained tests, and let $\pi(x)$ be the conditional probability of that recorded class. The retained source is $T_V \pi(v) = \int v\pi \, d\mu$, $v \in V$. Reweighting by w requires the same integral against wv, which may fall outside the retained span.

Theorem 2 (Intervention compatibility and sharp ambiguity)**.** *For bounded nonnegative w, its action on the compressed label source is well-defined for every π exactly when $wV \subseteq V$. For any required test $f \in L^1(\mu)$,*

$$\sup_{\substack{0 \leq \pi_+, \pi_- \leq 1 \\ T_V \pi_+ = T_V \pi_-}} \left| \int f(\pi_+ - \pi_-) \, d\mu \right| = \operatorname{dist}_{L^1(\mu)}(f, V). \tag{5}$$

The optimal worst-case prediction error from the retained information is half this distance. The same statements hold for each class coordinate of a multiclass source.

Two label distributions can agree on every retained statistic yet respond differently to a required test f; the distance gives their largest possible response difference. The proof uses Hahn–Banach duality; under invariance, the intervention is the adjoint of multiplication restricted to V. In the group

example, omitting the interaction's label moment can give opposite one-step changes in loss-based detection AUROC (area under the receiver operating characteristic curve). A 48-row construction demonstrates this ambiguity; retaining the moment resolves it (Section D).

A quantitative criterion for representation design. Suppose f is the label-weighted quantity a new intervention requires. Choose an approximating test $q \in V$. The retained source supplies $T_V\pi(q)$, while the known feature measure supplies a correction:

$$\widehat{m}_f = T_V\pi(q) + \frac{1}{2}\int (f-q)\,d\mu, \qquad \left|\widehat{m}_f - \int f\pi\,d\mu\right| \le \frac{1}{2}\|f-q\|_{L^1(\mu)}.$$

The formula follows by centring the unknown class probability at one half. Together with Theorem 2, it identifies the best worst-case accuracy available from this representation. On a finite feature support, finding q is weighted absolute-error approximation using the retained test columns. The intervention and feature geometry determine this calculation before its training outcome is observed. Adding an informative test enlarges V and reduces the remaining ambiguity. This is a quantitative way to decide what additional evidence a representation should retain. Training stability and the detector query then determine how source uncertainty affects the eventual ranking, as developed in Section 4. Our training implementation retains the required label responses exactly and controls approximation of the model-dependent feature term.

Dimensionality of the source representation. Feature measures belong to the dual of continuous functions on their compact support. Their action on training observables gives the relevant accuracy criterion. No finite-dimensional linear test space containing the constant can be invariant under multiplication by a weight with infinite essential range, since finite dimension would imply a polynomial identity for that weight. For an m-valued weight, at most m powers applied to each original test give a finite invariant enrichment. The required representation therefore depends on the permitted interventions. Fixed datasets and finite intervention horizons admit finite representations; controlled approximation supplies the economical version used below.

Interventions on neighbourhood propagation. In the bird example, a small neighbourhood may support the wrong annotation, whereas wider propagation may include informative bird examples. Predicting the effect of this change requires retaining the distribution of label evidence over distance. For exponential weighting, a source-weighted distance measure ν records the messages and a positive mass measure μ records the available weight. The response at inverse bandwidth t is

$$F(t) = \frac{\mathcal{L}\nu(t)}{\mathcal{L}\mu(t)}, \qquad \mathcal{L}\mu(t) = \int e^{-td}\,d\mu(d).$$

Larger t concentrates on nearby examples; smaller t admits distant ones. The entire response curve is determined by these measures, while the normalised query can identify an equivalence class of them.

Theorem 3 (Normalised propagation characterisation)**.** *For compactly supported finite signed source measures and nonzero positive mass measures, two normalised responses agree throughout an interval with nonempty interior if and only if*

$$\nu * \mu' = \nu' * \mu,$$

coordinatewise for vector sources, where $*$ denotes convolution.

Cross-multiplication turns equality of the ratios into equality of products of Laplace transforms; uniqueness then identifies the cross-convolutions. This cancellation occurs because spreading both message and mass profiles by the same independent distance-offset distribution multiplies both transforms by the same factor, leaving every normalised response unchanged. The theorem identifies exactly what this readout can distinguish. Theorems 1 and 3 therefore identify the information relevant to two response mechanisms under the permitted interventions, supporting approximation at the accuracy required for configuration selection.

4 From Characterisation to Economical Search

How to use the algorithm space. Specify the sources, interactions, readouts and proposed interventions. The characterisation then identifies the information needed to predict their effects, and

error bounds guide further computation. For the group example, retaining the intersection's label evidence resolves the source ambiguity; training and readout bounds connect the approximation to configuration selection.

Source reuse across candidate models. The gradient in Equation (4) repeatedly integrates the same data at different model states. Cached moments approximate softmax around each group's feature mean to first or cubic order, retaining the exact label term. Recomputing coefficients at each state preserves nonlinear candidate trajectories. Building on polynomial approximate statistics for generalised linear models (Huggins et al., 2017), we control source-action error for finite detector selection.

Forecast-based candidate screening. A training trajectory is a sequence of model states. Several configurations share that trajectory, differing in stopping time or detector readout. The screening policy ranks trajectories by their best predicted configuration, trains the four highest-ranked distinct trajectories, and selects among all their exact readouts. Its selection accuracy is tested prospectively.

Prediction error bounds. An error interval measures how far a forecast could be from the exact outcome. Carrying these intervals through selection additionally certifies the final choice, directing exact computation to unresolved comparisons. A fourth-order remainder for the cubic approximation gives a gradient-error bound $b_{a,t}$ for configuration a at step t. The global bound assumes the most sensitive response at every update. Intersecting it with a local bound using current class probabilities and feature spread sharpens the interval without changing the forecast, resolving more comparisons before exact evaluation. For learning rate γ_a, ridge λ_a, and $\kappa_a = \| \int w_a(x) x x^\top d\mu \|_{\mathrm{op}}$, define

$$\rho_a = \max\{|1 - \gamma_a \lambda_a|, |1 - \gamma_a(\lambda_a + \kappa_a/2)|\}.$$

The parameter-error radius obeys

$$E_{a,t+1} \le \rho_a E_{a,t} + \gamma_a b_{a,t}, \qquad E_{a,0} = 0. \tag{6}$$

Here, ρ_a determines how each update propagates existing error, while $\gamma_a b_{a,t}$ contributes new approximation error. When $\rho_a < 1$, earlier errors are attenuated; weaker contraction allows more error to accumulate, potentially requiring additional exact evaluations. Loss and competing-class margins are $\sqrt{2}\|x_i\|$-Lipschitz in W; AUM averages the corresponding margin radii along training. These score intervals determine which clean–noisy pairs have a resolved ranking and therefore give empirical AUROC bounds $[L_a, U_a]$ on a trusted development set. Proofs and the explicit local remainder appear in Sections B and C.

Theorem 4 (A reliable finite-family choice)**.** *For valid AUROC intervals over a finite candidate family, selecting $\widehat{a}$ when $\max_a U_a - L_{\widehat{a}} \le \tau$ guarantees $\max_a \mathrm{AUROC}(a) - \mathrm{AUROC}(\widehat{a}) \le \tau$. Exact evaluation of unresolved candidates preserves the guarantee and terminates after at most exhaustive evaluation of the family.*

The guarantee controls search accuracy on the specified development objective. Extending it to unreviewed examples requires statistical assumptions linking that objective to the target population, as with exhaustive selection.

Certified selection: predict, bound, refine.

1. Construct source statistics for the permitted interventions.
2. Predict trajectories and readouts; compute their AUROC intervals.
3. Discard candidates whose upper bounds cannot improve the incumbent beyond the chosen tolerance.
4. Refine unresolved candidates and stop when the interval gap is resolved.

A decision can be resolved before its scores are exact. Suppose an evaluated configuration achieves 85.0% AUROC. An alternative with interval $[84.2\%, 84.6\%]$ can be set aside. Another, predicted at 84.7% with interval $[84.1\%, 85.3\%]$, could still improve the choice. Tightening its interval to $[84.6\%, 84.8\%]$ settles the decision without changing the forecast or training that alternative. The matched bound experiment below tests this mechanism.

The required approximation accuracy depends on the separation between competing configurations. Coarse predictions can resolve clearly inferior candidates, while close comparisons require tighter bounds or exact evaluation. The characterisation thus guides both which information to retain and how accurately to represent it for the selection task.

5 Computational Cost and Selection Accuracy

Screening cost and accuracy. We first test whether forecasts reduce search cost while preserving selection accuracy. A prospective study covers 24 families on Dry Bean (Koklu & Ozkan, 2020), PenDigits (Alpaydin & Alimoglu, 1996) and all-class Food-101N (Lee et al., 2018), with two seeds and four regimes spanning wider interventions, weaker regularisation and longer horizons. Each combines 81 training trajectories with three stopping times and loss/AUM readouts, giving 486 choices and both beneficial and harmful interventions. First-order forecasts followed by four exact trajectories find an optimum in every family (Table 2). Each policy's cost includes forecasting, common preprocessing and selected-model execution. Successive halving retains earlier readouts and also finds every optimum. Screening takes less time than halving in all 24 families, corroborated by three-run timings on six predeclared families.

Policy	Max. regret (pp)	vs. exact	vs. halving
First-order forecast, one trajectory	0.00938	6.86×	2.43×
First-order forecast, four trajectories	0	5.49×	1.98×
Cubic certified selection	0	1.16×	0.49×
Successive halving	0	2.47×	1.00×

Table 2: **Reusable forecasts reduce search cost.** Median paired full-cost speed-ups over 24 primary families; values above one favour the row policy. Each primary timing is a single run. One-trajectory selection reranks its six exact readouts. The families share datasets and candidate structure. Complete controls and repeated timings appear in Section H.

Certified selection on a separate frozen family. We separately measure certification cost in 18 cases on the same datasets, using the same 486-choice structure and an interior reference. The training choices vary example weights, regularisation and learning rate. Fifteen cases vary seeds and dimensions; three Food-101N cases use 40,000 training examples. A predefined computational-size estimate selects cubic-moment approximation in fifteen cases and exact-prefix continuation in three. The latter runs initial training steps exactly and bounds the remaining trajectory. Both routes refine unresolved comparisons. Choices are saved before exhaustive auditing. All 18 attain zero development regret under the 0.1-pp tolerance, with no observed bound violations. Table 3 reports paired timings; full protocols appear in Section G.

Setting	Selection (s)	Exact grid (s)	Speed-up
Dry Bean, 8 components	0.898	1.341	1.50×
PenDigits, 8 components	1.245	1.307	1.05×
Food-101N, 6k / 8 components	6.282	24.925	3.98×
Food-101N, 6k / 16 components	16.674	26.916	1.61×
Food-101N, 6k / 32; prefix	29.366	30.581	1.04×
Food-101N, 40k / 16 components	24.811	169.549	6.83×

Table 3: **Certified selection with less computation.** Three-repeat paired CPU medians include source preparation, required verification and common loading and principal component analysis (PCA). The optimised exact grid shares trajectories and readouts and uses screened batching. Speed-ups are medians of paired ratios. All audited choices have zero development regret.

Paired timing repeats alternate method order. The large-Food speed-up ranges from 6.79× to 6.85×: source reuse is especially effective when many examples support a modest feature dimension.

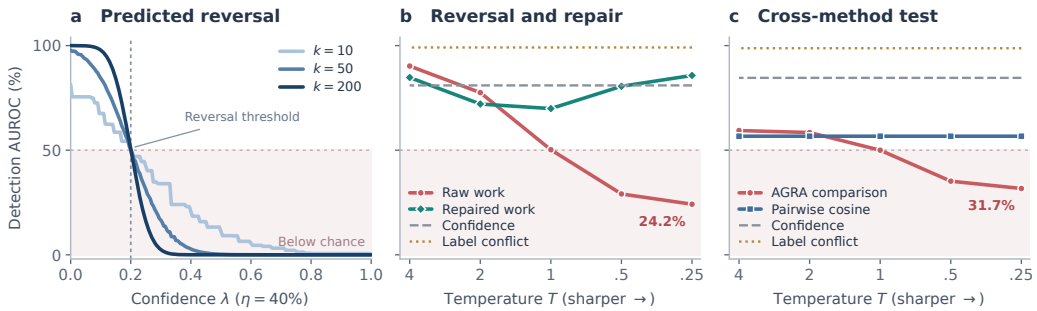


Figure 2: **Confidence can reverse agreement without changing predicted classes.** (a) Exact homogeneous-model AUROC, with reversal threshold $\lambda = 0.2$. (b) Frozen STL-10 (Coates et al., 2011) reversal and pilot repair. (c) AGRA comparison at epoch 30 reverses while pairwise cosine is invariant. Empirical curves average five class pairs and three seeds; lines connect tested temperatures. Direct label evidence remains strongest. Protocols and controls appear in Section M.

Sharper bounds reduce the cost of certification. To isolate the effect of the bounds, we hold cubic predictions fixed and vary only their error analysis. The probability-local remainder reduces additional exact trajectories from 30 to two on Food8 and from 23 to one on large Food16. In the latter matched control, selection cost falls from 78.98 to 21.23 seconds. A first-order source control needs 40 additional trajectories. All initially predict an optimum, so tighter bounds reduce verification time without changing the choice.

Certification accuracy and verification cost. Both approximation orders already select within 0.1 pp throughout the follow-up: certification corrects no consequential point-selection error. Its additional contribution is guaranteed error control, with a measured verification cost. Cubic certification is faster than exhaustive evaluation in 14 of 24 families and halving in four. Broad intervals can require extensive verification even when forecasts are accurate. The earlier stress tests likewise find six of eight certified searches slower than direct evaluation.

Selection across propagation choices. To test selection when neighbourhood weighting changes, a separate normalised-propagation study covers 45 frozen families. It increases polynomial degree before exactly evaluating unresolved candidates. Adaptive approximation meets the 0.1-pp target throughout: 44 choices have zero regret and the maximum is 0.0063 pp. Median speed-ups are $3.06\times$ for selection and $1.76\times$ including common preparation. Retained polynomial degrees vary across 2, 4, 8, 16 and 32. Its response characterisation therefore supports a second practical route to adaptive finite representation.

Held-out outcomes. We additionally evaluate whether configurations selected on the development set retain their advantage on separate examples from the same dataset. Twelve of the fifteen smaller cases improve over their reference and three decline; the largest decline is 1.54 pp. The three larger test changes are $+0.02$, -2.13 and -2.45 pp. The full audit reports every case, including gaps to the descriptive test optimum.

6 Predicting Confidence Reversal and Invariance

The algorithm space also supports predictions about individual design choices. Consider averaging individual gradient cosines or taking the cosine against a mean gradient. At a fixed binary network state, temperature rescales each nonzero gradient by a positive example-dependent factor. Normalising each source first removes those factors, making the pairwise comparison invariant. Aggregating raw sources retains them: changing confidence reweights the reference evidence and can reverse the comparison.

Predicted reversal and repair. For homogeneous binary references with independent flip rate $\eta < 1/2$ and true-class prediction probability $(1 + \lambda)/2$, define contextual work as $-r_i^\top \bar{r}_i$, where

$r_i = p_i - e_{\tilde{y}_i}$ and $\bar{r}_i$ is the mean reference residual. Its noisy-minus-clean gap is $1 - 2\eta - \lambda$. Sharpening confidence through this threshold reverses the query's interpretation while classifications remain unchanged. The analysis prescribes an orientation repair with the field fixed.

Tests of the mechanism. On frozen STL-10 (Coates et al., 2011), raw-work AUROC falls from 90.22% to 24.18%; the pilot repair restores 85.73%. A separate AGRA comparison test falls from 59.39% to 31.69% while pairwise cosine stays invariant (Section M.6). Reference composition also affects the comparison, as analysed in Section M.3. The design implication is to examine both the placement of normalisation and whether the readout's orientation remains appropriate after confidence changes (Figure 2).

7 Related Work and Scope

Unifying detector evidence. Model probing organises detectors by model, probe, ensemble and aggregation (George et al., 2024). Kernel smoothing, label propagation and sum-based set representations establish the aggregation lineage (Nadaraya, 1964; Zhu & Ghahramani, 2002; Zaheer et al., 2017). Class-based influence and AGRA study gradient comparisons (Nguyen-Duc et al., 2023; Sedova et al., 2023); our polarity analysis derives how normalisation placement controls confidence sensitivity.

From retained information to economical selection. ADGym learns pipeline rankings from historical datasets (Jiang et al., 2023); we derive forecasts and certification bounds from the current problem's training dynamics. The measure characterisation builds on sigmoid identification (Cybenko, 1989); classical optimal recovery supplies the duality behind sharp information ambiguity (Novak & Woźniakowski, 2008). Polynomial approximate sufficient statistics for generalised linear models (PASS-GLM; Huggins et al., 2017) support source approximation. Balcan et al. (2023) analyse approximate validation losses and regularisation tuning across problem instances. We connect intervention-specific information requirements to local multiclass response bounds, training trajectories and detector rankings, supporting economical screening and certified selection.

Scope and limitations. The training characterisation covers deterministic full-batch fixed-feature softmax models; necessity uses all initial states. The polarity guarantee assumes homogeneous binary noise and independent reference banks. Weak regularisation and long horizons can make certification expensive. Screening accuracy is empirical. Code includes numerical safeguards and bound audits. Timings cover configuration search. Measuring human-led discovery requires a separate study of scientific enquiry, whose subjective decisions vary across researchers and problems.

8 Conclusion

Our operator-theoretic characterisation provides a common representation of label-noise detectors through sources, interactions and queries. It specifies the information needed to predict how proposed changes affect detector behaviour. This account predicts when confidence reverses gradient agreement and when normalisation preserves its interpretation. It also identifies the information needed to forecast training trajectories and detector rankings. Reusing that information makes screening economical; propagating approximation bounds supports certified selection and quantifies its verification cost. Together, these results support detector design through predictions of intervention effects and allocation of computation to comparisons that can affect configuration selection. Future work will extend this account to jointly learned features, identify when certification prevents consequential errors at acceptable cost, and study when to repeat dataset-specific search or reuse configurations, linking intervention sensitivity to transfer.

Ethics Statement

Label-error scores are uncertain inferences. Rare, ambiguous, or minority examples can be correctly labelled despite local disagreement. Scores should support human review and subgroup analysis. Decisions about annotation quality or data removal require contextual assessment.

Reproducibility Statement

The characterisation proofs, approximation bounds and selection protocols appear in Sections A to H. Detector definitions and supporting experiments are detailed in Sections L and P to R. The implementation and experiment drivers are available at `https://anonymous.4open.science/r/label-noise-operator-164B/`. The repository's `README.md` gives setup and main experiment commands. Supporting studies are documented in `SUPPORTING.md`; `VALIDATION.md` records release checks.

AI Use Statement

Generative AI assisted with literature review, language editing, proof development and checking, source-code implementation and review, refinement, analysis, and verification of experimental protocols, all of which were also further cross-verified by the human authors.

References

Ethem Alpaydin and Fevzi Alimoglu. Pen-based recognition of handwritten digits. UCI Machine Learning Repository, 1996. URL `https://doi.org/10.24432/C5MG6K`.

Maria-Florina Balcan, Anh Nguyen, and Dravyansh Sharma. New bounds for hyperparameter tuning of regression problems across instances. In *Advances in Neural Information Processing Systems*, volume 36, pp. 80066–80078, 2023. URL `https://papers.neurips.cc/paper_files/paper/2023/file/fd62b65606f0f0d2af2c01623a224258-Paper-Conference.pdf`.

Ilies Chibane, Thomas George, Pierre Nodet, and Vincent Lemaire. Calibration improves detection of mislabeled examples, 2025. URL `https://arxiv.org/abs/2511.02738`. arXiv:2511.02738.

Adam Coates, Honglak Lee, and Andrew Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989. doi: 10.1007/BF02551274.

Thomas George, Pierre Nodet, Alexis Bondu, and Vincent Lemaire. Mislabeled examples detection viewed as probing machine learning models: concepts, survey and extensive benchmark. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=3YlOr7BHkx`.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Jonathan H. Huggins, Ryan P. Adams, and Tamara Broderick. PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://papers.nips.cc/paper_files/paper/2017/hash/07811dc6c422334ce36a09ff5cd6fe71-Abstract.html`.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Minqi Jiang, Chaochuan Hou, Ao Zheng, Songqiao Han, Hailiang Huang, Qingsong Wen, Xiyang Hu, and Yue Zhao. ADGym: Design choices for deep anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL `https://arxiv.org/abs/2309.15376`.

Suyeon Kim, Dongha Lee, SeongKu Kang, Sukang Chae, Sanghwan Jang, and Hwanjo Yu. Learning discriminative dynamics with label corruption for noisy label detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

Murat Koklu and Ilker Ali Ozkan. Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, 174:105507, 2020. doi: 10.1016/j.compag.2020.105507. URL https://archive.ics.uci.edu/dataset/602/dry+bean+dataset.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. CleanNet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456, 2018. URL https://kuanghuei.github.io/Food-101N/.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, 2023.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020.

Thang Nguyen-Duc, Hoang Thanh-Tung, Quan Hung Tran, Dang Huu-Tien, Hieu Nguyen, Anh T. V. Dau, and Nghi Bui. Class based influence functions for error detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1204–1218, 2023. URL https://aclanthology.org/2023.acl-short.104/.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Erich Novak and Henryk Woźniakowski. *Tractability of Multivariate Problems, Volume I: Linear Information*, volume 6 of *EMS Tracts in Mathematics*. European Mathematical Society, 2008. doi: 10.4171/026.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, 2021.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, 2020.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, 2020.

Anastasiia Sedova, Lena Zellinger, and Benjamin Roth. Learning with noisy labels by adaptive gradient-based outlier removal. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pp. 237–253, 2023. URL https://arxiv.org/abs/2306.04502.

Natasha Singh-Miller and Michael Collins. Learning label embeddings for nearest-neighbor multi-class classification with an application to speech recognition. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://www.cs.columbia.edu/~mcollins/papers/nips09.pdf.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A. Osborne. On the limitations of representing functions on sets. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6487–6494, 2019.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.12088.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Haoran Zhang, Aparna Balagopalan, Nassim Oufattole, Hyewon Jeong, Yan Wu, Jiacheng Zhu, and Marzyeh Ghassemi. LEMoN: Label error detection using multimodal neighbors. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74447–74489, 2025.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, 2021.

Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*, 2022.

A SOURCES SUFFICIENT FOR INTERVENTION PREDICTION

The representation must preserve the information needed to determine an intervention's effect. For fixed-feature softmax training, this information has a concrete form: feature measures describe how predictions are integrated, while label moments supply the observed-label contribution. The intervention family determines which combinations of these sources matter.

Setting. Let z range over a compact subset $K \subset \mathbb{R}^d$ and let $x(z) = (z, c_0)$, where the constant intercept $c_0 \neq 0$ is known. There are $C \geq 2$ classes. Fix a common objective normalisation and two sign-valued directions h_1, h_2. The four joint sign groups have finite positive feature measures μ_g and label moments

$$B_g = \int x(z) e_y^{\mathsf{T}} \, d\mu_g(z, y).$$

The feature marginal of the labelled measure is understood in μ_g. Intervention (a, b) assigns weight $w_g(a, b) = 1 + ah_{1,g} + bh_{2,g} \geq 0$. Define three channels

$$(\mu^{(j)}, B^{(j)}) = \sum_g h_{j,g}(\mu_g, B_g), \qquad h_{0,g} = 1, \quad j = 0, 1, 2.$$

Channels one and two may be signed. The unregularised gradient field is

$$G_{a,b}(W) = \sum_{j=0}^{2} \alpha_j \left\{ \int x \, \mathrm{softmax}(W^{\mathsf{T}} x)^{\mathsf{T}} \, d\mu^{(j)} - B^{(j)} \right\}, \qquad \alpha = (1, a, b).$$

Ridge adds the known term λW. Each candidate takes deterministic, full-batch updates with its declared fixed learning rate and ridge strength.

Functional representation and intervention operators. By the Riesz representation theorem, finite signed source measures form $C(K)^*$ with the total-variation norm. The three-channel source space is

$$\mathcal{S} = \left(C(K)^* \times \mathbb{R}^{(d+1)\times C} \right)^3.$$

With summed component norms, the weight intervention is the bounded linear map $\mathcal{L}_\alpha(s_0, s_1, s_2) = \sum_j \alpha_j s_j$, of norm at most $\|\alpha\|_\infty$. At each fixed W, integrating the bounded continuous matrix function $x \, \mathrm{softmax}(W^{\mathsf{T}} x)^{\mathsf{T}}$ and subtracting the label moment is another bounded linear map of the source. Training composes these maps with the evolving W, producing nonlinear parameter dynamics; the retained detector query then reads the resulting trajectory. Thus linear source operators coexist with nonlinear evolution and readout. The declared family reweights the original source once. Arbitrarily composing weight changes by multiplication may introduce the additional $h_1 h_2$ channel and requires enriching the source space.

The finite representation below approximates how the source acts on the observables needed by training. For a bounded parameter region, the relevant accuracy is expressed by the query seminorm

$$\|\delta\mu\|_{\mathcal{Q}_R} = \sup_{\substack{\|W\|_F \leq R \\ \|V\|_F \leq 1}} \left| \int \left\langle V, x \, \mathrm{softmax}(W^{\mathsf{T}} x)^{\mathsf{T}} \right\rangle_F \, d\delta\mu \right|.$$

Adding $\|\delta B\|_F$ bounds the corresponding gradient-action error. The moment construction need not itself define an approximate measure; its stepwise defect bounds directly control this query-restricted action at each visited parameter state. The ambient dual space gives the exact characterisation, while the observables actually used determine the approximation accuracy needed for selection.

We first establish a three-channel formulation, then derive Theorem 1 for general known weight families below.

Theorem 5 (Three-channel training-map characterisation)**.** *Consider two source states satisfying the setting above. Suppose the admissible coefficient vectors* $(1, a, b)$ *span* $\mathbb{R}^3$*. The following are equivalent:*

1. *their three channel pairs* $(\mu^{(j)}, B^{(j)})$ *agree;*

2. *their gradient fields $G_{a,b}(W)$ agree for every admissible intervention and every initial parameter W;*

3. *their one-step training maps agree for those interventions and every initial W, with common known regularisation and nonzero learning rates.*

Consequently the channel pairs determine all finite training trajectories under the declared interventions. With a fixed evaluation/query object Q, they also determine the corresponding detector outputs.

Proof. Linearity in α and the spanning assumption recover each channel gradient from the intervention fields. Consider one channel, allowing its measure to be finite signed, and suppose two channel gradients agree. Send the intercept logit of class r to $-\infty$ while keeping other logits zero. Its probability converges uniformly to zero. Compactness and finite total variation imply that the gradient column converges to $-B_r$, identifying every label-moment column.

After subtracting the equal label moments, read the intercept coordinate of class r. Assign its logit $u^{\mathsf{T}}z + v + \log(C-1)$ and all other logits zero. Dividing that coordinate by c_0 gives equality of the integrals of $\mathrm{sigmoid}(u^{\mathsf{T}}z + v)$ for all u, v. Sending v to $+\infty$ identifies total mass. Replacing (u, v) by $(tu, -ts)$ and taking $t \to \infty$ identifies every projected half-line measure at thresholds that are not atoms. A finite signed measure has at most countably many atoms; continuity points determine its one-dimensional projection measure. Equality of all projections, equivalently of characteristic functions in all directions, identifies the feature measure. This proves necessity. The converse is substitution. Subtracting the common initial state and dividing by the nonzero step size gives the update-map equivalence; the known ridge term cancels. Iteration proves trajectory sufficiency, and applying the retained query proves detector-output sufficiency. □

For known deterministic feature groups G_g, identifying $\mu^{(0)}$ also identifies every $\mu_g = \mathbf{1}_{G_g}\mu^{(0)}$, including the h_1h_2 feature channel. The three feature channels are therefore redundant on this restricted source class. The group label moments can retain an unidentified contrast $\delta B_g = h_{1,g}h_{2,g}D$, provided these changes are realisable by their label distributions. Four groups remain convenient for local approximation. The necessity statement uses all initial parameters. The retained query Q includes evaluation features, recorded labels and the required history/readout definition; together with the training map, these determine the per-example scores.

Proof of Theorem 1. Let $\mathcal{H}$ be a linear space of known bounded Borel weight functions containing 1, and let ν be the joint feature–label measure with feature marginal μ. Define the bounded label-response operator

$$\mathcal{B}(h) = \int h(x)xe_y^{\mathsf{T}}\, d\nu(x, y), \qquad \|\mathcal{B}(h)\|_F \le \|h\|_\infty \int \|x\|\, d\mu.$$

Then equality, for every W and every $h \in \mathcal{H}$, of

$$G_h(W) = \int h(x)x\,\mathrm{softmax}(W^{\mathsf{T}}x)^{\mathsf{T}}\, d\mu - \mathcal{B}(h)$$

is equivalent to equality of μ and $\mathcal{B}|_{\mathcal{H}}$. Indeed, $h = 1$ and the single-channel argument in the proof of Theorem 5 identify μ; subtracting its known weighted softmax term then identifies $\mathcal{B}(h)$ for each h. Conversely, substitution gives every gradient field. It suffices to observe nonnegative admissible weights whose linear span is $\mathcal{H}$. For the implemented family, $\mathcal{H} = \mathrm{span}\{1, h_1, h_2\}$, so one feature measure and three label-moment matrices suffice. Enlarging $\mathcal{H}$ enriches the label-response operator; no second independent feature measure is needed. This statement assumes the weight functions are fixed and known. Deterministic thresholds derived from the common μ also agree when their tie convention is fixed.

This description also specifies when richer interventions require richer label information. Let the feature coordinates be independent and uniform on $\{-2, -1, 1, 2\}$, with intercept 1, and assign the function u the values $(-1, 2, -2, 1)$ on these four points. Set $h_j = \mathrm{sign}(x_j)$ and $\delta(x) = u(x_1)u(x_2)/4$. Two binary-labelled source states with the same feature measure have class-one probabilities $1/2 \pm \varepsilon\delta(x)$, where $0 < \varepsilon \le 1/2$. Since

$$\mathbb{E}u = \mathbb{E}[x_j u] = 0, \qquad \mathbb{E}[h_j u] = -1/2,$$

their label-response operators agree on $1, h_1, h_2$, including every feature coordinate. Their complete training maps therefore agree throughout the original intervention family. But $\mathbb{E}[h_1 h_2 \delta] = 1/16$, so the class-one intercept entry of $\mathcal{B}(h_1 h_2)$ differs by $\varepsilon/8$. Adding this interaction weight distinguishes the training maps. If weights must be nonnegative, use $1 + \gamma h_1 h_2$ with $0 < |\gamma| \leq 1$. The difference arises from label-response information that the original representation does not retain, even though the feature geometry is identical and fully known.

At the other extreme, suppose $\mathcal{H} \subset C(K)$ is uniformly dense in $C(K)$. The intercept coordinates of $\mathcal{B}(h)$ give the integrals of h against each class measure, up to the fixed nonzero intercept. Boundedness extends equality from $\mathcal{H}$ to $C(K)$, and uniqueness in the Riesz representation theorem identifies every class measure, hence the full joint labelled measure. In particular, a unital separating algebra of continuous weight functions is sufficient by Stone–Weierstrass. Thus the intervention family determines exactly which label observables must be retained. Multiplying continuous reweightings can generate such an algebra; repeated multiplication of two sign functions alone instead closes on the four-dimensional span of $1, h_1, h_2, h_1 h_2$.

Why the complete source can be infinite-dimensional. The feature measures belong to a finite product of Banach spaces of signed measures with total-variation norms. Their equality is characterised by the training maps above. Suppose the allowed support contains a nondegenerate line segment within a fixed intervention group. For any degree $m \geq 1$, choose $m + 2$ distinct scalar positions on that segment. Their degree-m Vandermonde matrix has a nonzero null vector. Its positive and negative parts have equal nonzero mass and matching moments through degree m; normalising them gives distinct positive measures with those same moments. Uniform class proportions also give matching label moments. The theorem then implies a distinction in some training map despite the equal truncated representation. Thus no universal fixed moment order is exact over these arbitrary compact source measures. A fixed finite support can be recovered from sufficiently many moments, and a finite empirical measure is itself an exact finite representation. Practical compression concerns controlled approximation.

B NONLINEAR SOURCE INTEGRATION WITH A FINITE CERTIFICATE

Within a nonempty group write $x = c + u$, where c is its mean and the intercept coordinate of u is zero. At the current candidate parameter W, let

$$p = \operatorname{softmax}(W^{\mathsf{T}} c), \qquad v_r = W_r - \sum_s p_s W_s.$$

Define the integrated contractions

$$J_k(v_1, \ldots, v_k) = \mathbb{E}_g \left[x \prod_{j=1}^{k} (u^{\mathsf{T}} v_j) \right].$$

Here $J_0 = c$. The cubic source approximation in column r is

$$\begin{aligned} \widehat{S}_r(W) = p_r \Big\{ & J_0 + J_1(v_r) \\ & + \frac{1}{2} \big[J_2(v_r, v_r) - \sum_s p_s J_2(v_s, v_s) \big] \\ & + \frac{1}{6} \big[J_3(v_r, v_r, v_r) - 3 \sum_s p_s J_3(v_r, v_s, v_s) \big] \\ & - \frac{1}{6} \sum_s p_s J_3(v_s, v_s, v_s) \Big\}. \end{aligned}$$

It is integrated exactly from centred feature moments through degree four. Subtract the exact label moment and add the exact ridge term. The group-centre probabilities are recomputed at the current candidate state, so this source approximation retains nonlinear training dynamics.

Lemma 1 (Global and local fourth-derivative bounds). *Let* $p(t) = \mathrm{softmax}(z + ta)$ *and* $\bar{a} = a - C^{-1}(\mathbf{1}^\mathsf{T} a)\mathbf{1}$. *Uniformly in* z, t, C,

$$\|p^{(4)}(t)\|_1 \le \frac{14}{3}\|\bar{a}\|_2^4.$$

Writing $p_{\max} = \max_r p_r(t)$ *also gives*

$$\|p^{(4)}(t)\|_2 \le \sqrt{2}(2p_{\max} + 10p_{\max}^2)\|\bar{a}\|_2^4.$$

Proof. Centre $A = a_Y - \mathbb{E}_p a_Y$ and put $\mu_k = \mathbb{E}_p A^k$. Differentiation gives $p_r^{(4)} = p_r(A_r^4 - 6A_r^2\mu_2 - 4A_r\mu_3 + 6\mu_2^2 - \mu_4)$. If $L = \max a - \min a$, endpoint convexity bounds $\mu_2 \le L^2/4$, $\mu_4 \le L^4/12$, and $\mathbb{E}|A|^3 \le L^3/8$. Therefore

$$\|p^{(4)}\|_1 \le 2\mu_4 + 12\mu_2^2 + 4\sqrt{\mu_2}\,\mathbb{E}|A|^3 \le (7/6)L^4 \le (14/3)\|\bar{a}\|_2^4.$$

For the local bound, test against a unit vector b and centre $B = b_Y - \mathbb{E}_p b_Y$. The derivative pairing is $\mathbb{E}[BA^4] - 6\mathbb{E}[BA^2]\mu_2 - 4\mathbb{E}[BA]\mu_3$. Use $\mathrm{range}(b) \le \sqrt{2}$, $\mathrm{range}(a) \le \sqrt{2}\|\bar{a}\|$, $\mu_2 \le p_{\max}\|\bar{a}\|^2$, and $|\mathrm{Cov}(a, b)| \le p_{\max}\|\bar{a}\|$. The three terms are bounded respectively by $2\sqrt{2}p_{\max}\|\bar{a}\|^4$, $6\sqrt{2}p_{\max}^2\|\bar{a}\|^4$, and $4\sqrt{2}p_{\max}^2\|\bar{a}\|^4$. Duality completes the proof. □

Proposition 1 (Computable integrated-source error). *Let* $H = I - \mathbf{1}\mathbf{1}^\mathsf{T}/C$, $M = WHW^\mathsf{T}$, *and*

$$E_g(W) = \mathbb{E}_g[\|x\|_2(u^\mathsf{T} M u)^2].$$

The group gradient error is at most $(7/36)E_g(W)$. *Let* r_g *bound* $\|u\|$ *and define*

$$p_{g,\max} = \max_r \mathrm{softmax}(W^\mathsf{T} c)_r, \qquad k_g = \sqrt{2}r_g\|W_f H\|_{\mathrm{op}}.$$

The error is also at most

$$\sqrt{2}\left[2p_{g,\max}\phi_4(k_g) + 10p_{g,\max}^2\phi_4(2k_g)\right] E_g(W),$$

Here

$$\phi_4(k) = \frac{e^k - 1 - k - k^2/2 - k^3/6}{k^4}, \quad \phi_4(0) = \frac{1}{24}.$$

Both bounds can be evaluated from the current state and a cached positive fourth moment. Here W_f *contains only the non-intercept feature rows: the intercept coordinate of* u *is zero, so this restricted operator norm suffices. Nonnegative group weights combine their errors additively.*

Proof. The cubic Taylor remainder integrates $p^{(4)}(s)$ against $(1 - s)^3/6$. The global bound follows from Lemma 1 and $\|xy^\mathsf{T}\|_F \le \|x\|\|y\|_1$. Along the local logit segment, $p_{\max}(s) \le p_{g,\max}e^{sk_g}$. Integrating the local bound yields $\phi_4(k_g)$ and $\phi_4(2k_g)$. Finally E_g is the contraction of $\mathbb{E}_g[\|x\|u^{\otimes 4}]$ with M twice. The weighted fourth moment must be retained; it need not be recoverable from the unweighted fourth moment. □

C Adaptive computation with a common selection guarantee

For a candidate let $K = \|\mathbb{E}[wxx^\mathsf{T}]\|_{\mathrm{op}}$. The true regularised softmax Hessian lies between λI and $(\lambda + K/2)I$. Its gradient update is therefore ρ-Lipschitz, where

$$\rho = \max\{|1 - \eta\lambda|, |1 - \eta(\lambda + K/2)|\}.$$

Proposition 2 (Approximate-source and exact-prefix certificates). *If the source approximation at* $\widehat{W}_t$ *has gradient error at most* b_t, *then*

$$E_{t+1} = \rho E_t + \eta b_t, \qquad E_0 = 0, \quad \Longrightarrow \quad \|W_t - \widehat{W}_t\|_F \le E_t.$$

Alternatively, from an exact prefix state W_t *with gradient* $g(W_t)$,

$$\|W_{t+k} - W_t\|_F \le \eta\|g(W_t)\|_F \sum_{j=0}^{k-1} \rho^j.$$

Proof. For the first bound, add and subtract the exact update at $\widehat{W}_t$ and apply its Lipschitz constant plus the source defect. For the second, consecutive exact increments contract by at most ρ; sum their norms. The finite sums are valid for any finite ρ; the implementation's exact-prefix branch requires $\rho < 1$ for useful continuation bounds. □

Endpoint cross-entropy and maximum competing margin have Lipschitz constant $\sqrt{2}\|x_i\|$ in W. Their state-error radii therefore give detector-score intervals; average margins use the average radius. For an exact prefix, retain its observed margins and bound each future margin using the second inequality above. The fixed evaluation labels and query definitions are part of Q. Trusted corruption statuses are used only to evaluate the empirical detection objective.

Proof of Theorem 4. Noisy–clean pair contributions are monotone in the noisy score and antitone in the clean score. Comparing lower noisy endpoints to upper clean endpoints gives L_a; the opposite comparison gives U_a. Hence $L_a \le \mathrm{AUROC}(a) \le U_a$, so

$$\max_a \mathrm{AUROC}(a) - \mathrm{AUROC}(\widehat{a}) \le \max_a U_a - L_{\widehat{a}} \le \tau.$$

Representations may be chosen from source moments or exact prefixes and refined using unresolved intervals, provided every retained interval remains valid. Adaptive computation and intersection preserve these deterministic bounds. Exact evaluation collapses a candidate's interval to its AUROC; evaluating each unresolved candidate at most once therefore terminates after at most exhaustive evaluation of the finite family. □

The production gate uses a fixed computational size proxy to choose between the two certified engines. The selection result applies to the declared finite family and trusted development objective, with fixed features and deterministic full-batch training. The analytical bounds hold in exact arithmetic; the implementation uses finite-input checks and numerical padding. Experimental protocols and numerical audits appear in Section G.

D When an intervention acts on a compressed label source

An intervention can be applied to a compressed source only when the retained information determines its result. The following characterisation makes that requirement exact, and measures the ambiguity when it fails.

Fix a finite positive feature measure μ. Binary labels are represented by measurable conditional probabilities $\pi \in L^\infty(\mu)$ with $0 \le \pi \le 1$. Let V be a closed real linear subspace of $L^1(\mu)$. The retained source is the functional

$$T_V\pi : v \longmapsto \int v\pi \, d\mu, \qquad v \in V.$$

For the training-source characterisation, take $V = \overline{\mathrm{span}\{x_j h : h \in \mathcal{H},\ j = 1, \dots, d+1\}}^{L^1(\mu)}$. Features are bounded in the compact-support setting, and the nonzero constant intercept with $1 \in \mathcal{H}$ ensures $1 \in V$.

For a known nonnegative weight $w \in L^\infty(\mu)$, the compressed intervention in Theorem 2 is the transformation

$$T_V\pi \longmapsto \left(v \longmapsto \int wv\pi \, d\mu \right)$$

whose well-definedness we characterise below.

Sharpness uses the full class of conditional probabilities $0 \le \pi \le 1$. If each distinct empirical feature atom must instead carry a single hard label, the admissible class is discrete: the same distance is a valid upper bound on ambiguity, but need not be attainable. Replicated feature atoms can realise fractional conditional probabilities with suitable denominators.

Proof of Theorem 2. We first establish the sharp ambiguity formula in Equation (5). Differences of two admissible probabilities have the form $\delta = \pi_+ - \pi_-$ with $\|\delta\|_\infty \le 1$. Conversely every such δ is

realised by $\pi_\pm = (1 \pm \delta)/2$. Equality of the retained sources is equivalent to $\delta \in V^\perp$, where

$$V^\perp = \{\delta \in L^\infty(\mu) : \int v\delta \, d\mu = 0 \text{ for every } v \in V\}.$$

For any $v \in V$, the absolute target difference is at most $\|f - v\|_1$. This gives the upper bound. The Hahn–Banach theorem applied to the quotient $L^1(\mu)/V$ supplies a norm-one annihilating functional attaining the quotient norm of $f + V$. The dual identification $(L^1(\mu))^* = L^\infty(\mu)$ represents it by such a δ, proving the reverse inequality and attainment. The constructed probabilities belong to the central fibre $T_V\pi = T_V(1/2)$ because δ annihilates V. This is a worst-case fibre diameter; it need not be attained in every fibre.

For descent, take $f = wv$ for every retained test v. Its ambiguity vanishes exactly when $wv \in V$, since V is closed. Equivalently, under invariance, the compressed intervention is the adjoint of the restricted multiplication operator $M_w : V \to V$. Its operator norm is at most $\|w\|_\infty$.

For $C \geq 2$ classes, retain $(T_V\pi_c)_{c=1}^C$ with the conditional probability vector π in the simplex. The same condition $wV \subset V$ characterises descent: sufficiency holds coordinate by coordinate, and necessity follows by embedding the binary witness into two classes. For any fixed class coordinate, the exact ambiguity of $\int f\pi_c \, d\mu$ remains $\mathrm{dist}_1(f, V)$, with minimax error half that value; another class absorbs the complementary probability. An arbitrary joint vector-valued query requires its own norm-dependent analysis.

The output above is an unnormalised weighted label source, whose class mass need not be a probability relative to the original μ. If the feature measure is also reweighted to $w\mu$, the same integrals describe its label source. Normalisation by a positive known total feature mass does not change the identifiability criterion. A fixed feature measure is essential to the statement: it isolates missing label information from uncertainty in geometry.

A quantitative test for representation enrichment. For a desired retained query v, the distance $\mathrm{dist}_1(wv, V)$ is the largest possible change in its weighted label response between indistinguishable sources. Any prediction using only the compressed source incurs worst-case absolute error at least half this diameter. Conversely, choosing a retained test $q \in V$ and using

$$\int q\pi \, d\mu + \tfrac{1}{2} \int (f - q) \, d\mu$$

to estimate $\int f\pi \, d\mu$ gives error at most $\|f - q\|_1/2$. Taking the infimum shows the exact minimax error is $\mathrm{dist}_1(f, V)/2$ (an infimum over estimators suffices if a best L^1 approximant is not attained). The offset uses the known feature measure. These are label-source ambiguities; a detector-performance consequence still requires the training and readout stability analysis. □

A sharp finite example. Let μ be uniform on $\{-2, -1, 1, 2\}^2$, let $h_j = \mathrm{sign}(x_j)$, and retain $V = \mathrm{span}\{x_j h : j = 1, 2, 3,\ h \in \{1, h_1, h_2\}\}$, where $x_3 = 1$. For $f = h_1h_2$, the exact ambiguity is $1/6$. Indeed, the retained approximant $q = (x_1h_2 + x_2h_1)/3$ gives

$$f - q = h_1h_2\{1 - (|x_1| + |x_2|)/3\}, \qquad \|f - q\|_1 = 1/6.$$

For the matching lower bound, write $x_1 = s_1a$, $x_2 = s_2b$ with $a, b \in \{1, 2\}$, and set $\delta = s_1s_2t_{ab}$, where $t_{11} = 1$, $t_{12} = t_{21} = 1/3$, and $t_{22} = -1$. Sign symmetry makes all retained pairings vanish except potentially those against x_1h_2 and x_2h_1. These also vanish because $\mathbb{E}[at_{ab}] = \mathbb{E}[bt_{ab}] = 0$. Thus $\delta \in V^\perp$, $\|\delta\|_\infty = 1$, and $\int f\delta \, d\mu = \mathbb{E}[t_{ab}] = 1/6$. The indistinguishable probabilities $(1 \pm \delta)/2$ attain that diameter; the exact worst-case prediction error is $1/12$. Here the conditional probabilities are in $\{0, 1/3, 2/3, 1\}$, so both source states have exact deterministic-label realisations on 48 rows: three copies of each of the 16 feature points, with the corresponding number of class-one labels. The two states share the same replicated feature data. For a nonnegative intervention $w = 1 + \gamma h_1h_2$, $|\gamma| \leq 1$, the target $w \cdot 1$ has ambiguity $|\gamma|/6$, since $1 \in V$.

The same construction shows that the omitted information can determine the sign of a detector-performance change. Use the two probabilities just constructed, initialise binary softmax regression at $W = 0$, and use any nonnegative weight from the original span $\{1, h_1, h_2\}$. Both source states have zero gradient at the origin and remain there under gradient descent with ridge. Their original

training trajectories and detector outputs agree. Under the new positive weight $w = 1 + \gamma h_1 h_2$, $0 < \gamma < 1$, symmetry gives

$$\mathbb{E}[wx_1\delta] = \mathbb{E}[wx_2\delta] = 0, \qquad \mathbb{E}[w\delta] = \gamma/6.$$

After one step of size $\eta > 0$, the two source states therefore have class intercepts

$$(\eta\gamma/12, -\eta\gamma/12) \quad \text{and} \quad (-\eta\gamma/12, \eta\gamma/12).$$

Fix a common two-example evaluation query at the feature vector $(0, 0, 1)$, with recorded labels class one and class two respectively, and with the first example clean and the second noisy. Loss-based detection has AUROC 50% before intervention, 100% in the first source state afterwards, and 0% in the second. Thus an unchanged compressed source permits opposite signs of the intervention effect, although enriching it with the interaction label moment resolves the ambiguity. This worst-case construction starts from tied scores and illustrates why ranking margins matter when translating source uncertainty to detector uncertainty. The evaluation query and its clean/noisy statuses are held fixed across source states.

Corollary 1 (Repeated interventions and representation size)**.** *Suppose $1 \in V$ and V is finite-dimensional. If $wV \subset V$, then w has finite essential range. Thus a bounded weight with infinite essential range admits no finite-dimensional exact linear test representation containing the constant and closed under arbitrary repetitions of that intervention.*

Conversely, if w takes m distinct values almost everywhere and the initial test space has a finite basis $v_1, \dots, v_r$, then

$$V_w = \operatorname{span}\{w^k v_j : 0 \le k < m,\ 1 \le j \le r\}$$

is a finite-dimensional invariant enrichment, of dimension at most mr.

Proof. Invariance and $1 \in V$ imply $1, w, w^2, \dots \in V$. Finite dimension gives a nonzero real polynomial P with $P(w) = 0$ almost everywhere. Its finitely many real roots contain the essential range of w. Conversely, the polynomial whose roots are the m attained values expresses w^m as a linear combination of its lower powers; multiplying by each v_j proves invariance of V_w. Finite-dimensional subspaces of L^1 are closed. □

The infinite-dimensional obstruction concerns arbitrarily many repetitions. For a fixed maximum of K repetitions and a finite initial test space V_0, the finite space $\sum_{k=0}^{K} w^k V_0$ retains all required responses through that horizon; it need not be invariant under further multiplication.

For two sign controls, repeated products close on $\{1, h_1, h_2, h_1 h_2\}$. Including their products with the feature coordinates therefore gives a finite invariant label-source representation. Continuous weights with infinitely many essential values can require an unbounded sequence of new exact tests, even though approximations with controlled action error remain finite. On a fixed finite empirical support every weight has finite essential range, so this obstruction concerns universal or non-atomic source classes; a fixed dataset has an exact finite representation.

Mathematical provenance. The annihilator distance formula uses classical Hahn–Banach duality and optimal recovery; the descent condition uses multiplication invariance. Together they specify the intervention-dependent label observables that an exact linear training-source representation must retain.

E A LOWER-ORDER SOURCE CONTROL

The first-order control asks how much source detail is needed to obtain a useful certificate. It retains the same exact label contribution and expands the probability source to first order within each feature group. At a group centre c, put $p = \operatorname{softmax}(W^{\mathsf{T}} c)$ and $J_p = \operatorname{diag}(p) - pp^{\mathsf{T}}$. Its integrated source is

$$\widehat{S}_g^{(1)}(W) = cp^{\mathsf{T}} + \mathbb{E}_g[xu^{\mathsf{T}}] W J_p.$$

Recomputing p at each step retains nonlinear dependence on the current model.

Lemma 2 (First-order source certificate). *Let* $a_0 = a - C^{-1}(\mathbf{1}^\mathsf{T} a)\mathbf{1}$ *and* $P = \max_r \operatorname{softmax}(z)_r$. *Then*

$$\|D^2 \operatorname{softmax}(z)[a,a]\|_2 \le \min\{1/\sqrt{2}, \sqrt{2}P\}\|a_0\|_2^2.$$

Consequently, using the group notation of Proposition 1, a valid integrated source error is

$$\min\left\{\frac{1}{2\sqrt{2}}, \sqrt{2}p_{g,\max}\phi_2(k_g)\right\}\mathbb{E}_g[\|x\|_2 u^\mathsf{T} W H W^\mathsf{T} u], \qquad \phi_2(k) = \frac{e^k - 1 - k}{k^2}, \quad \phi_2(0) = \frac{1}{2}.$$

The weighted second moment makes this bound inexpensive to evaluate.

Proof. For a unit test vector b, centre the class-indexed variables $A = a_Y - \mathbb{E}_p a_Y$ and $B = b_Y - \mathbb{E}_p b_Y$. The second-derivative pairing is $\mathbb{E}_p[BA^2]$. Since $|B| \le \operatorname{range}(b) \le \sqrt{2}$, its absolute value is at most $\sqrt{2}\operatorname{Var}_p(a)$. The variance is bounded by both $P\|a_0\|^2$ and $\operatorname{range}(a)^2/4 \le \|a_0\|^2/2$. Taking the supremum over b proves the derivative bound. Along a within-group logit segment, $P_s \le p_{g,\max} e^{s k_g}$. Integrating against the first-order Taylor remainder weight $(1-s)$ gives $\phi_2(k_g)$; integrating the uniform bound gives $1/(2\sqrt{2})$. Multiplication by $\|x\|$ and integration over the group complete the proof. □

These bounds use the same finite-training and selection results as the cubic source. A lower-order source saves preparation, while a larger remainder can require more exact verification. The reported order controls measure this trade-off; the frozen production study uses its declared cubic/prefix routing rule and adapts the exact verification work to the unresolved comparisons.

F Characterising normalised propagation responses

Changing a bandwidth changes the relative contributions of neighbouring examples. To predict this response, the source representation retains their distances together with their evidence. For one query, let μ_0 be a nonzero finite positive measure on $[0, R]$, and let μ_r be signed source-coordinate measures satisfying $|\mu_r| \le M_r \mu_0$. On data, these are $\mu_0 = \sum_j w_j \delta_{d_j}$ and $\mu_r = \sum_j w_j s_{jr} \delta_{d_j}$, with $w_j > 0$. Write

$$\mathcal{L}_\mu(\beta) = \int e^{-\beta d}\, d\mu(d), \qquad F_r(\beta) = \frac{\mathcal{L}_{\mu_r}(\beta)}{\mathcal{L}_{\mu_0}(\beta)}.$$

The vector measures form the Banach space $\mathcal{M}([0,R])^{p+1}$ with summed total-variation norms. Bandwidth increments act by bounded multiplication operators $J_\delta \mu = e^{-\delta d}\mu$, with $J_\delta J_\eta = J_{\delta+\eta}$ and $\|J_\delta\| \le e^{|\delta|R}$. Integrating the transformed measures and normalising produces the field; the detector query maps this field to a score. The ratio is locally Lipschitz where its positive denominator is bounded below.

Proof of Theorem 3. Consider two systems (μ_0, μ_r) and (ν_0, ν_r). The proof requires only finite signed source measures on compact support and nonzero positive mass measures; the domination bounds above are used in the response-error estimates below. Positive denominators allow cross-multiplication. The difference of the resulting products is the Laplace transform of $\eta_r = \mu_r * \nu_0 - \nu_r * \mu_0$. Compact support makes this transform entire. The identity theorem makes it identically zero when it vanishes on an interval with nonempty interior. All moments of η_r therefore vanish. Polynomial density in continuous functions on the compact support, followed by uniqueness of finite signed measures, gives $\eta_r = 0$. Conversely, the convolution identity equates the transform products, and division by positive denominators gives equality at every real bandwidth. The same argument applies whenever the fields agree on any real bandwidth set with a finite accumulation point, extending the interval condition. □

For a restricted linear query family $\mathcal{Q}$, replace the coordinate criterion by $(\sum_r q_r \mu_r) * \nu_0 = (\sum_r q_r \nu_r) * \mu_0$ for $q \in \mathcal{Q}$. Coordinate-separating queries recover the full criterion. A single maximum-contrast score may identify fewer distinctions. Common convolution by a positive measure cancels from numerator and denominator, providing an intuitive example of response equivalence. Tilting respects convolution, so bandwidth increments act consistently on these equivalence classes.

Proposition 3 (Fixed-order moments need not determine bandwidth responses)**.** *For every $m \geq 0$ and $R > 0$, two valid two-class source systems have identical moments through degree m and distinct normalised responses at every $\beta > 0$.*

Proof. Set $n = m + 1$, $h = R/n$ and $\sigma = 2^{-n} \sum_{j=0}^{n} (-1)^j \binom{n}{j} \delta_{jh}$, $\lambda = |\sigma|$. Finite differences give zero moments of σ through degree m. Use the common denominator measure λ and class measures $(\lambda + \sigma)/2, (\lambda - \sigma)/2$, then swap the classes for the second system. These measures are nonnegative and sum to λ. Their first-class response difference is

$$\frac{\mathcal{L}_\sigma(\beta)}{\mathcal{L}_\lambda(\beta)} = \left(\frac{1 - e^{-\beta h}}{1 + e^{-\beta h}}\right)^n = \tanh(\beta h/2)^n > 0.$$

□

The example establishes the need for increasing representational richness across arbitrary source measures. Its shrinking difference also explains why finite approximations can be accurate on a bounded intervention range.

Computable response intervals. At anchor b, centre distances by $t = d - c$, $|t| \leq r$, and let P_b be the probability measure proportional to $e^{-bd}\mu_0$. For $\delta = \beta - b$ and source density $s = d\mu_r/d\mu_0$, write $D = \mathbb{E}_b e^{-\delta t}$ and $N = \mathbb{E}_b s e^{-\delta t}$. Their degree-$m$ Taylor approximations D_m, N_m satisfy

$$e_D = \frac{e^{|\delta| r} |\delta|^{m+1}}{(m+1)!} \mathbb{E}_b |t|^{m+1}, \qquad e_N = \frac{e^{|\delta| r} |\delta|^{m+1}}{(m+1)!} \mathbb{E}_b |s||t|^{m+1}.$$

Taylor's integral remainder gives $|D - D_m| \leq e_D$ and $|N - N_m| \leq e_N$. Whenever $D_m - e_D > 0$, subtraction of the ratios yields

$$\left| F - \widehat{F} \right| \leq \frac{e_N + |\widehat{F}| e_D}{D_m - e_D}, \qquad \widehat{F} = N_m / D_m.$$

Clipping to the known source range preserves the bound. An unresolved denominator uses the unconditional range bound and triggers refinement. For interpolation at distinct nodes $\delta_0, \ldots, \delta_m \in [-a, a]$ and a target $\delta \in [-a, a]$, replace $e^{|\delta| r} |\delta|^{m+1}$ by $e^{ar} |\prod_j (\delta - \delta_j)|$; the derivative form of the scalar interpolation remainder proves the same ratio guarantee.

A tighter remainder from numerator–denominator cancellation. For $0 \leq s \leq 1$ and an unmodified ratio $\widehat{F} \in [0, 1]$, apply the remainder to $\mathbb{E}_b[(s - \widehat{F}) e^{-\delta t}]$, whose polynomial approximation vanishes at the target. The inequality $|s - \widehat{F}| \leq (1 - \widehat{F}) s + \widehat{F}(1 - s)$ replaces the numerator above by the smaller expression

$$A\{\widehat{F} A_D + (1 - 2\widehat{F}) A_N\}, \qquad A_D = \frac{\mathbb{E}_b |t|^{m+1}}{(m+1)!}, \quad A_N = \frac{\mathbb{E}_b s |t|^{m+1}}{(m+1)!},$$

where A is the appropriate exponential or interpolation factor. If the raw ratio is outside the source range, use the preceding general bound.

Construction itself can exploit positivity. For nonnegative sources, let $a_j = \mathbb{E}_b[s t^j]/j!$. For odd k, Cauchy–Schwarz gives

$$\frac{\mathbb{E}_b[s|t|^k]}{k!} \leq \sqrt{\frac{k+1}{k} a_{k-1} a_{k+1}}.$$

Even orders are already absolute moments. Thus two extra polynomial orders supply the needed remainder moments with a shared source contraction. The same construction applies to the denominator using $s = 1$.

A fixed linear query has error at most $\sum_r |q_r| e_r$; a maximum competing-class contrast has error at most $\max_c e_c + e_{\widetilde{y}}$. These score intervals enter Theorem 4. Increasing the polynomial order and exactly evaluating unresolved candidates therefore preserves the same selection guarantee. The tested engine uses orders 2, 4, 8, 16 and 32, followed by exact fallback.

G INTERVENTION-SELECTION PROTOCOLS AND COMPLETE EVIDENCE

The study measures the computation required to choose among a declared family of detector–training configurations. Each evaluated candidate shares its trajectory across stopping horizons and readouts.

G.1 TRAINING DATA AND INTERVENTION FAMILY

Dry Bean (Koklu & Ozkan, 2020) and PenDigits (Alpaydin & Alimoglu, 1996) use 1,800 training, 400 development and 400 test examples. Training-only standardisation and principal components retain eight coordinates, divided by the maximum training norm; an intercept of 0.3 is appended. Feature-dependent corruption has probability $15\% + 35\%/(1 + \exp(-8x_1))$; replacements are uniform among other classes. Seeds 1001–1003 determine partitions and corruption. Food-101N (Lee et al., 2018) uses naturally corrupted recorded labels and all 101 classes, with cached ResNet-18 features. The smaller protocol uses 6,000/800/800 rows and 8, 16 or 32 components; the larger uses 40,000/1,200/2,000 rows and 16 components, with seeds 1101–1103. SHA-256 image grouping removes exact duplicates before splitting. PCA and scaling are fitted on training rows. Verified annotations supply development/test statuses; training uses recorded labels. The cases share three datasets.

The controls h_1, h_2 are signs of the first two feature coordinates relative to their training medians. We vary $w = 1 + ah_1 + bh_2$, with $a, b \in \{-0.3, 0, 0.3\}$, ridge $\lambda \in \{0.015, 0.03, 0.06\}$ and step size $\eta \in \{3, 4, 5\}$. The 81 deterministic full-batch softmax trajectories start at zero. Horizons 20, 40 and 60 are read using loss and negative area under the margin, producing 486 choices. The objective divides weighted cross-entropy by the training sample count; ridge regularises every parameter, including the intercept. Food features also use intercept 0.3 and maximum-training-norm scaling. The smaller Food loader charges its initial 32-component PCA before retaining the requested coordinates and rescaling them on training rows. The fixed reference is unweighted, with ridge 0.03, step size 4, horizon 40 and AUM.

The source route retains exact group label moments and a cubic probability approximation with the weighted fourth-moment remainder in Proposition 1. Production uses source moments when $4\binom{d+2}{3} < n_{\text{train}}$ and exact-prefix continuation otherwise. This cubic-term cost proxy routes the smaller Food32 cases to prefixes. Each route refines unresolved candidates until the empirical AUROC gap is at most 0.1 percentage points (pp). Construction, forecasts and required exact evaluations all enter selection cost.

G.2 FULL FROZEN RESULTS

The audit contains 15 source cases and three prefix cases. Source forecast matrices and required verification results precede exhaustive evaluation. The held-out audit hashes each existing choice before evaluating test outcomes; no selection is updated. The table retains every outcome.

G.3 COST ACCOUNTING AND CONTROLS

The timed implementation uses single-thread CPU NumPy on an AMD EPYC 74F3 virtual machine with 18 available vCPUs. Set `OPENBLAS_NUM_THREADS=1` and `OMP_NUM_THREADS=1`. The environment is Python 3.12.3, NumPy 2.5.2, SciPy 1.17.1 and PyTorch 2.7.1+cu126; PyTorch loads cached features. The optimised exhaustive comparator caches class–feature totals, uses stable in-place softmax and screens batch sizes 1, 3 and 9. It shares 60 updates across six horizon/readout choices. Three quiet paired timing repeats alternate selector/comparator order. Both receive the same loading/PCA charge within each pair. Cached feature extraction is excluded equally. The separate exhaustive scientific audit is excluded from deployment cost. Ratios are medians of within-pair ratios.

Attributing savings to the bound. The global/local ablation fixes every cubic point forecast. Additional exact trajectories fall from 11 to 6 on Dry Bean, 46 to 33 on PenDigits, 30 to 2 on Food8, 30 to 1 on Food16, and 23 to 1 on large Food16, using the first seed in each group. Food8 selection-only cost falls from 12.756 to 4.309 seconds; large Food16 falls from 78.978 to 21.233 seconds. All five controls pass forecast-identity and interval-nesting checks. These mechanism-ablation times are single runs; the main speed-up table uses the separate paired protocol.

Dataset	d	Seed	Engine	Final gap (pp)	Extra exact	Dev. AUROC	Test AUROC	Test Δ (pp)	Test oracle gap (pp)
Dry Bean	8	1001	S	0.0000	6	93.03	92.43	13.127	0.000
Dry Bean	8	1002	S	0.0867	13	92.42	92.91	9.314	1.178
Dry Bean	8	1003	S	0.0089	8	94.11	94.16	11.159	0.384
Food, 6k	16	1001	S	0.0480	1	78.69	76.84	1.665	0.000
Food, 6k	16	1002	S	0.0276	1	70.62	72.10	-1.540	1.603
Food, 6k	16	1003	S	0.0000	1	72.93	75.05	2.035	0.002
Food, 6k	32	1001	P	0.0000	–	79.37	77.92	1.122	0.425
Food, 6k	32	1002	P	0.0000	–	72.04	73.07	-1.464	1.515
Food, 6k	32	1003	P	0.0000	–	74.02	75.44	2.096	1.072
Food, 6k	8	1001	S	0.0550	2	74.34	73.51	-0.102	0.120
Food, 6k	8	1002	S	0.0265	1	68.04	72.61	0.063	0.000
Food, 6k	8	1003	S	0.0346	1	72.10	71.26	1.310	0.019
PenDigits	8	1001	S	0.0000	33	94.04	94.43	6.080	0.198
PenDigits	8	1002	S	0.0000	27	92.00	93.33	4.021	0.000
PenDigits	8	1003	S	0.0000	27	94.41	92.85	2.172	1.607
Food, 40k	16	1101	S	0.0326	1	74.61	76.49	0.015	0.000
Food, 40k	16	1102	S	0.0125	1	76.98	74.94	-2.134	2.155
Food, 40k	16	1103	S	0.0053	1	78.95	74.00	-2.445	2.463

Table 4: All 18 frozen training cases. S: source moments; P: exact prefixes. Every row has zero observed development regret and zero reported AUROC-bound violations. Source parameter-bound audits also pass. AUROC columns are percentages. Final gap is the certified selection uncertainty. Extra exact counts trajectories beyond the source route's initial reference; the prefix route refines partial trajectories. Test Δ is relative to the fixed reference. The test oracle is computed descriptively after selection. Values are rounded only for display.

Approximation order and empirical search. The large first-order control needs 40 additional trajectories and 113.065 seconds, compared with one trajectory and 21.233 seconds for the local cubic source. Both initially predict an optimum. Point-only forecasts also select optima in checked cases; including execution of the chosen configuration, the large-case selection cost is 17.032 seconds. Optimised successive halving shares paths through 81, 27 and 9 survivors at horizons 20, 40 and 60, totalling 2,340 updates. Its three-repeat selection-only medians are 11.117 seconds on Food8, 12.005 on Food16 and 80.300 on large Food16, with zero observed regret. Source selection is faster on Food8 and large Food16; halving is faster on Smaller Food16. These controls measure the computation required to attach an error guarantee to a successful forecast.

Economic stress tests. Eight development cases combine Dry Bean or PenDigits, first-order or cubic sources, and horizons $(20, 40, 60)$ or $(20, 100, 200)$, with ridge strengths $\{0.001, 0.005, 0.03\}$ and seeds 901 and 902. All have zero observed regret and no reported bound violations; six take longer than their direct comparator. First-order/cubic additional exact counts are 56/42 and 54/54 for the two Dry Bean horizons, and 80/54 and 80/54 for PenDigits. Longer trajectories and weaker regularisation can make verification nearly exhaustive even when the point forecast already identifies an optimum. These exploratory timings establish the economic boundary observed in the study.

G.4 Complementary propagation study

Five data/noise settings (Dry Bean and PenDigits with symmetric or instance-dependent corruption, and CIFAR-10N with human labels (Wei et al., 2022)), seeds 87–89, and bandwidth ranges $[0, 3]$, $[0, 12]$, $[0, 48]$ give 45 families. Each range has 65 bandwidths. The source representation uses evidence from 6,000 training examples, with 256 development and 512 test examples. A training-standardised logistic model (regularisation parameter $C = 1$, the limited-memory Broyden–Fletcher–Goldfarb–Shanno optimiser, at most 300 iterations) supplies class probabilities. Sources concatenate recorded-label one-hot vectors and these probabilities. Squared Euclidean distances are divided by the maximum distance from the first 1,000 training examples to the training population. At each bandwidth, write the two propagated fields as α and $\bar{p}$. For $a \in \{0, 1/4, 1/2, 3/4, 1\}$, set $f = \alpha - a\bar{p}$ and evaluate its maximum competing-class contrast and the five queries $(bp - e_{\widetilde{y}})^{\mathsf{T}} f$ with $b \in \{0, 1/4, 1/2, 3/4, 1\}$. These 30 readouts give 1,950 choices per family. Tabular symmetric corruption is 40%; instance-dependent corruption is $15\% + 50\%/(1 + e^{-z})$, where z centres the first raw coordinate by its sample median and scales by its standard deviation. CIFAR-10N uses its human worse-label annotation and cached ImageNet ResNet-18 features. The positive-moment engine refines

through orders 2, 4, 8, 16 and 32, then exactly evaluates unresolved candidates. Controls include fixed-order moments, certified kernel interpolation and point-only interpolation. All 45 adaptive choices meet the 0.1-pp tolerance; 44 have zero observed regret and the maximum is 0.00629 pp. Median conditional and full speed-ups are 3.06 and 1.76. Exact fallback uses at most three bandwidth evaluations. Full cost includes common representation preparation. Adaptive selection reduces selection time by approximately 7.96% relative to fixed degree 32. Point-only degree-8 interpolation is faster and also has zero observed regret. This study is reported separately from the 18 training cases.

G.5 Code, artefacts and mathematical provenance

The repository's `README.md` gives preparation and launch commands. From the repository root, `reproduce.py` provides the `certified` and `large-food` stages. For direct invocation, from `experiments/`, run `certified_source_selection.py` with seeds 1001–1003 and `source_selection_large_food.py` with seeds 1101–1103, using new output directories. All paths here are relative to the released repository. Exact and paired timing entry points are `source_controls_batched_exact.py` and `source_selection_paired_timing.py`. Controls and audits use `source_controls_batched_halving.py`, `source_controls_point_moments.py`, `linear_source_moment_probe.py` and `source_controls_test_transfer.py`. The propagation entry point is `run_economical_response_benchmark.py`.

Frozen protocols and predictions recorded SHA-256 digests before their oracle audits. The original validation covered source-byte and artefact checks, 30 regression tests and analytical bound audits. The primary dependency supplement is explicitly post-validation. Digests identify bytes; the ordered pipeline documents prediction, selection and evaluation. Fresh runs generate new records in the chosen output directories; they do not establish the original chronology. The supplied test runner and `VALIDATION.md` describe release validation separately from the historical checks above.

Measure identification builds on affine-sigmoid discrimination (Cybenko, 1989); polynomial approximate sufficient statistics have an established role in generalised linear models (Huggins et al., 2017). Sharp ambiguity uses optimal recovery from linear information (Novak & Woźniakowski, 2008). Approximate validation losses and regularisation tuning across instances are studied by Balcan et al. (2023). Here these tools yield an intervention-dependent characterisation of detector evidence, finite-trajectory error control, and a measured reduction in the exact computation needed for reliable selection.

H Prospective comparison of forecasting, screening and certification

The characterisation supplies reusable evidence for predicting candidate trajectories. This follow-up separates the practical value of those forecasts from the additional cost and guarantee of certification. Its protocol fixes both approximation orders, screening policies and timing repetitions before the primary evaluations. It is separate from the 18-case study in Section G.

H.1 Data, interventions and selection rules

Dry Bean and PenDigits use 1,800/400/400 training/development/test examples; all-101-class Food-101N uses 6,000/800/800. The loaders, training-only preprocessing, corruption and duplicate controls follow Section G, with eight retained components and new seeds 1201 and 1202. Four regimes on each dataset give 24 primary families. Both first-order and cubic approximations are tested on each, giving 48 order-specific cells. The development tolerance is 0.1 pp.

The reference uses zero weight coefficients, the middle ridge, step size four and middle-horizon AUM. Every family contains beneficial and harmful interventions relative to this reference; the best available improvements range from 0.51 to 12.95 AUROC pp. Test outcomes do not guide selection.

Regime	Each weight coefficient	Ridge	Horizons
A	$-0.3, 0, 0.3$	$0.015, 0.03, 0.06$	$20, 40, 60$
B	$-0.45, 0, 0.45$	$0.015, 0.03, 0.06$	$20, 40, 60$
C	$-0.45, 0, 0.45$	$0.001, 0.005, 0.03$	$20, 100, 200$
D	$-0.45, 0, 0.45$	$0.0001, 0.001, 0.005$	$20, 200, 600$

Table 5: Predeclared follow-up families. Every regime uses step sizes $3, 4, 5$, weights $w = 1 + ah_1 + bh_2$, and loss and negative-AUM readouts: 81 trajectories and 486 choices. C and D jointly vary regularisation and horizon.

Point forecasts and screening. For each approximation order, point-only selection omits error moments, bounds and nominal exact training. It executes the predicted-best trajectory and reranks its six exact horizon/readout choices. We also retain regret for the stricter forecast-selected readout. The predeclared screening policy executes the four distinct trajectories with the highest predicted scores and selects among their exact readouts. Each policy's cost includes fresh source construction, forecasting and selected-model execution. The fixed four-trajectory rule supplies an empirical screening policy; certified selection uses the interval-based stopping rule in Theorem 4.

Certification and budget controls. Raw certification evaluates the reference and predicted-best candidate, then prioritises unresolved upper bounds, breaking ties by point forecasts. Exact trajectories and their readouts are cached within each policy. A separate cost-aware hybrid budgets source construction, forecasting and interval computation at 20% of an estimated remaining exhaustive cost: 80 times the measured reference-trajectory cost. On budget exhaustion it finishes unresolved candidates exactly. It may also complete directly when every unresolved upper bound exceeds the incumbent by more than tolerance. All sunk work is charged. This routing rule is a measured-cost heuristic.

Comparators and timing. Optimised successive halving retains $81 \to 27 \to 9$ trajectories at the three horizons and keeps earlier readouts eligible. Exhaustive evaluation shares trajectories across horizons and readouts. All methods use the same exact backend, with batch sizes nine for tabular data and one for Food, fixed from the earlier timing screen. Common loading/PCA and selected-model execution enter every total; cached feature extraction is excluded equally. The later exhaustive scientific audit and diagnostic bound checks are excluded from deployment cost. Each method is charged for all outcomes it computes or uses.

Family order is shuffled with seed 21201 and method order rotates deterministically. Each primary family receives one timing run. Two further runs are predeclared for A and D at seed 1201 on each dataset, giving six three-run families and 36 runs overall. Speed-ups are paired full-cost ratios; reported medians are taken over those ratios.

H.2 Prediction quality and incremental certification

First-order and cubic point policies attain the exact optimum in 22 and 23 of 24 families respectively; both meet the 0.1-pp target throughout. The strict forecast-selected readouts also meet that target. Both four-trajectory policies attain the optimum in all 24 families. Median per-family forecasting mean absolute error (MAE) is 0.04555 pp for first order and 0.000659 pp for cubic approximation. Median Spearman agreement across the 486 choices is 0.999784 and 0.999997, respectively. Thus more accurate score forecasts need not change the selected configuration.

Table 2 reports the primary practical comparison. The remaining controls appear below, avoiding duplication of those rows. All certified and hybrid choices have zero observed regret. Neither order produces a point-selection error above 0.1 pp, so there are zero consequential corrections and zero economical consequential corrections across the 48 order-specific cells. The predeclared latter endpoint required both a corrected error above tolerance and a certified cost below exhaustive evaluation. Broad intervals can require many exact trajectories despite accurate point predictions. Both hybrid orders have median cost above exhaustive evaluation.

Additional policy	Max. regret (pp)	vs. exact	vs. halving	Faster than halving
Cubic, one trajectory	0.00865	3.716×	1.518×	20/24
Cubic, four trajectories	0	3.357×	1.385×	19/24
First-order certificate	0	0.965×	0.379×	0/24
First-order hybrid	0	0.855×	0.383×	0/24
Cubic hybrid	0	0.831×	0.405×	4/24

Table 6: Additional predeclared controls: median paired full-cost speed-ups over the 24 initial runs. All methods meet the 0.1-pp tolerance.

Family	Point / halving	Four / halving	Four / exact	Cubic cert. / halving
Dry Bean, A	1.747	1.604	2.744	0.808
Dry Bean, D	1.485	1.215	5.306	0.158
PenDigits, A	2.704	2.258	4.534	0.581
PenDigits, D	1.714	1.323	6.255	0.149
Food-101N, A	3.924	3.125	5.941	2.259
Food-101N, D	4.164	2.441	11.339	0.199

Table 7: Three-run median paired speed-ups at seed 1201. Point and Four use first-order forecasts; values above one favour the forecast-based policy. Four-trajectory screening reaches an optimum in every run.

H.3 Repeated timings and interpretation

First-order screening remains faster than halving in all six repeated families. The studies reuse datasets, seeds and candidate structure, so their counts are descriptive. Optima concentrate in a few configurations, and halving also selects an optimum throughout. The demonstrated practical gain is computation saved on these structured, fixed-feature searches. The source forecast is the shared algorithmic contribution across the point, screening and certified policies; evaluating a fixed shortlist is a familiar screening principle. Certification supplies a prescribed error guarantee whose observed cost depends on approximation tightness.

H.4 Implementation, provenance and complete records

The anonymous review materials include `certification_followup.py`, `certification_followup_forecast.py` and `certification_followup_exact.py` as the runner, source predictor and exact comparator implementations. The protocol is `CERTIFICATION_FOLLOWUP_PROTOCOL.md`; the report generator is `certification_followup_report.py`, all under `experiments/`. The root `README.md` documents the `search` stage of `reproduce.py`. The machine-readable report records every primary family and degree, strict and reranked regret, forecast agreement, runtime components, exact trajectory counts, initial interval widths, final gaps and paired repeats.

Protocol and transitive source digests were saved before primary evaluation, forecasts before selected execution or refinement, and choices before the exhaustive audit. Local hashes establish integrity rather than independent trusted timestamps. An initial provenance-collector error occurred before data loading and was corrected before restarting the study. After 35 runs, an external interruption stopped the last timing repetition. Its partial files were preserved and hashed; the complete family was rerun with its original method order after checking the frozen sources and matching CPU configuration. Partial timings were kept separate. Neither event changed the selection rules or primary outcomes.

All 36 scheduled runs completed. Validation verified 21 frozen source digests, 770 saved JSON digests and 22 archived partial files; all 12 regression tests passed. No score-bound or parameter-bound violation was observed. The 2,047 unavailable hybrid forecasts were intentionally abandoned at the declared budget cutoff; no numerical divergence was recorded. In the raw timing records, verification time includes reference time, which is also reported separately as a subset.

I Latent Labels and the Oracle Query

Let $(X, Y^\star, \widetilde{Y})$ take values in $\mathcal{X} \times [C] \times [C]$. The clean conditional-label field is

$$\eta_c(x) = \mathbb{P}(Y^\star = c \mid X = x), \tag{7}$$

and the instance-dependent transition field is

$$T_{cr}(x) = \mathbb{P}(\widetilde{Y} = r \mid Y^\star = c, X = x). \tag{8}$$

We call $D = \mathbf{1}\{Y^\star \neq \widetilde{Y}\}$ the corruption indicator. Conditioning Bayes' rule on an observed label r gives the oracle corruption query

$$q_X(x, r) = 1 - \frac{\eta_r(x) T_{rr}(x)}{\sum_{c=1}^{C} \eta_c(x) T_{cr}(x)}. \tag{9}$$

The denominator is the observed-label field $\mathbb{P}(\widetilde{Y} = r \mid X = x)$. Thus an oracle needs both clean-label compatibility and the corruption mechanism; disagreement with a modal nearby label is generally insufficient.

For a detector observing a representation $Z = \phi(X)$ rather than X, define

$$q_Z(z, r) = \mathbb{P}(D = 1 \mid Z = z, \widetilde{Y} = r).$$

Let $R(g) = \mathbb{P}(g(U) \neq D)$ denote corruption-classification risk using the information U available to g.

Proposition 4 (Perfect detection and irreducible ambiguity)**.** *The Bayes detector $g^\star(U) = \mathbf{1}\{q^\star(U) \geq 1/2\}$ has risk*

$$R^\star = \mathbb{E} \min\{q^\star(U), 1 - q^\star(U)\}. \tag{10}$$

Perfect individual corruption detection is possible if and only if $q^\star(U) \in \{0, 1\}$ almost surely.

This condition is stronger than having a smooth or clusterable clean-label field. Genuine class overlap, representation collisions, and locally plausible corruptions all produce nonzero Bayes risk. Moreover, $T(x)$ is not identifiable in general from one noisy label per instance (Liu et al., 2023); anchor points, class-conditional structure, multiple labels, clusterability, or other assumptions supply information that the observation model otherwise lacks (Zhu et al., 2021).

Proposition 5 (Representation cannot improve Bayes risk)**.** *If Z is a deterministic representation of X, or more generally a randomised representation satisfying $D \perp Z \mid (X, \widetilde{Y})$, then*

$$\mathbb{E} \min\{q_Z, 1 - q_Z\} \geq \mathbb{E} \min\{q_X, 1 - q_X\}. \tag{11}$$

Furthermore,

$$\mathbb{E}|q_X - q_Z| \leq \sqrt{\frac{1}{2} I(D; X \mid Z, \widetilde{Y})}. \tag{12}$$

Hence $I(D; X \mid Z, \widetilde{Y}) = 0$ is a sufficient condition for the representation to preserve the oracle corruption posterior.

The information condition concerns preservation of the corruption posterior. A representation may support accurate clean-label classification while omitting instance attributes that govern annotation mistakes; conversely, variation irrelevant to clean classification may help identify the corruption process.

J Field-Query Operators

Definition 1 (Admissible field-query detector)**.** Conditioned on a permitted training state, an admissible detector specifies:

1. coordinates $z_i = \phi(x_i)$;
2. example-local sources $r_i = \rho(x_i, \widetilde{y}_i)$ in a vector space $\mathcal{V}$, possibly containing a recorded training trajectory;

3. a bounded pairwise linear propagator $K(z_i, z_j) : \mathcal{V} \to \mathcal{W}$;
4. a compatibility map $C(\widetilde{y}_i, r_i, F_i)$; and
5. a calibration map a.

For a specified nonzero normalisation Z_i, its leave-one-out field, query, and score are

$$F_i = \frac{1}{Z_i} \sum_{j \neq i} K(z_i, z_j) r_j, \qquad Q_i = C(\widetilde{y}_i, r_i, F_i), \qquad \widehat{q}_i = a(Q_i). \tag{13}$$

Scalar neighbourhood kernels conventionally use $Z_i = \sum_{j \neq i} K(z_i, z_j)$; signed or matrix-valued interaction fields may instead use a fixed normalisation such as n. Time integration is allowed. The construction must remain permutation invariant and may not encode clean labels or D_i.

The definition covers explicit neighbourhood fields directly. Learned predictors fit through an implicit or dynamic version: training aggregates example-local sources into parameters, after which predictions query that compressed field. Section K makes this representation exact for instantaneous neural-network training.

J.1 Representation and its exact scope

An unrestricted fitted state could encode the detector itself. We therefore specify that state explicitly and characterise the evidence aggregation stage. Fix a state s, a query u, and a finite multiset M of context observations (z, c), where $c \in [C]$ is the recorded label. State and coordinates remain fixed when M is perturbed. Let $E_s(u, M) \in \mathbb{R}^d$ be evidence before compatibility. Normalised averages use separate numerator and mass channels, with division performed by the query.

Theorem 6 (Additive evidence representation and source sufficiency)**.** *Assume $E_s(u, \varnothing) = 0$. Then multiset additivity $E_s(u, M \uplus N) = E_s(u, M) + E_s(u, N)$ is equivalent to the existence of a unique single-observation response $g_s(u, z, c)$ satisfying*

$$E_s(u, M) = \sum_{(z,c) \in M} g_s(u, z, c). \tag{14}$$

Write $A_{s,u,z} = [g_s(u, z, 1) \ \cdots \ g_s(u, z, C)]$. For a specified local source matrix $R_{s,z} = [\rho_s(z, 1) \ \cdots \ \rho_s(z, C)] \in \mathbb{R}^{m \times C}$, a factorisation $A_{s,u,z} = K_s(u, z) R_{s,z}$ exists for every u if and only if

$$\ker R_{s,z} \subseteq \bigcap_u \ker A_{s,u,z}. \tag{15}$$

The minimum source dimension at z is

$$m_s^\star(z) = \dim \operatorname{span}\{\textit{rows of } A_{s,u,z} : u \in \mathcal{U}\} \leq C. \tag{16}$$

One-hot sources always suffice within this additive class. Uniformly bounded single-observation responses give uniformly bounded one-hot propagators.

The characterisation applies to the specified additive aggregation class. A detector additionally applies a declared query and calibration to evidence and query-local information. Training and neighbour selection must either admit explicit stagewise representations or be supplied as preprocessing. Top-k selection can change when observations are inserted; its graph is fixed only for the subsequent aggregation stage. General set-function representation results have their own dimensional and regularity requirements (Wagstaff et al., 2019).

Corollary 2 (Unrecoverable source information)**.** *If $v \in \ker R_{s,z}$ but $A_{s,u,z} v \neq 0$, no linear propagator of those sources can realise the requested evidence. Every candidate obeys*

$$\|A_{s,u,z} - K_s(u, z) R_{s,z}\|_{\mathrm{op}} \geq \frac{\|A_{s,u,z} v\|}{\|v\|}. \tag{17}$$

If two whole contexts give identical aggregate sources but their target scores differ by δ, every query of those sources has maximum error at least $\delta/2$ on the two contexts.

At uniform probabilities, every label has the same residual norm. Such a scalar source cannot express evidence distinguishing those labels; vector residuals retain the contrasts. This supplies a precise reason to test directional sources, without guaranteeing improved ranking. Similarly, if a local graph transmits no messages from outside a corrupted cluster, changes outside that cluster leave every query of its field unchanged. Addressing this obstruction requires additional information or a change of propagation.

Scope and regularity. Additivity is imposed at the evidence-aggregation stage; nonlinear queries of counts can therefore express nonlinear set statistics. The minimum source dimension is defined for the specified evidence map, and may differ for another map that produces the same ranking. Coordinate collisions must share a response map or require richer coordinates. For a compressed source on an infinite coordinate domain, pointwise kernel containment alone does not ensure a uniformly bounded propagator; a uniformly bounded factorisation is an additional requirement.

Source dimension for a family of two-label queries. Suppose each query compares only two classes, but all queries must reuse a shared linear label encoder R. Let G be the graph whose edges are the requested class pairs and let B_G have rows $e_b^\top - e_a^\top$, one per distinct edge. A scalar answer per query need not admit a scalar shared source.

Corollary 3 (Shared-source dimension for a family of label contrasts)**.** *For the prescribed evidence family B_G, the minimum dimension of a shared linear source admitting exact linear readout is $C - c(G)$, where $c(G)$ counts connected components, including isolated classes. In particular, connected pairwise query families require $C - 1$ source channels. For $B_G \neq 0$, any m-channel encoder and linear readout K satisfy*

$$\frac{\|B_G - KR\|_F}{\|B_G\|_F} \geq \frac{(\sum_{j>m} \sigma_j(B_G)^2)^{1/2}}{\|B_G\|_F}.$$

Equality is attained by retaining the first m right singular directions.

Proof. The kernel of B_G consists of vectors constant on each connected component, so its rank is $C - c(G)$. Apply Theorem 6 to the joint row span. Every KR has rank at most m; the optimal rank-m approximation retains the first m singular directions and leaves the stated squared tail error. □

The consequence is about reusable evidence: a different scalar source could answer each separate comparison, but cannot generally replace one shared source serving the whole family. A connected path of class comparisons already forces $C - 1$ channels. The graph-rank identity is standard; the design implication is obtained by applying source sufficiency across queries. The bound concerns exact prescribed contrasts and linear shared encoding, not arbitrary nonlinear codes, query-dependent sources, or equivalent rankings.

Cached multiclass audit. We examine all four reconstructed disjoint CIFAR-100 bundles. Each query compares its recorded class to the highest-probability alternative under the saved out-of-fold head. The resulting graphs have 2,161–3,275 distinct edges and are all connected: their minimum shared dimension is 99 despite each query requesting only one contrast. A 99-channel basis reproduces the contrast matrix to maximum error 3.11×10^{-15}. Even the optimal 64-channel linear encoding has relative Frobenius error 49.15%–53.57%. This is error on the equally weighted distinct contrast rows, not an AUROC loss or an empirical-frequency-weighted score error. The audit excludes aggressive lossless compression of this query family; the one-channel saving itself is modest. It uses no clean labels and is a retrospective algebraic audit, not a new detector or prospective forecast. `characterization_payoff_audit.py` also checks connected paths, disconnected pairs, and complete graphs.

Proposition 6 (Trusted-label selection bound)**.** *Fix a trusted reference bank and L candidate score functions independently of a fresh evaluation sample. Conditional on having $n_+ > 0$ corrupted and $n_- > 0$ clean independent evaluation examples, let $\widehat{A}_\ell$ and A_ℓ be empirical and population AUROC, with half credit for ties. With probability at least $1 - \delta$, simultaneously for all candidates,*

$$|\widehat{A}_\ell - A_\ell| \leq \epsilon = \sqrt{\frac{1}{2}\left(\frac{1}{n_+} + \frac{1}{n_-}\right)\log\frac{2L}{\delta}}.$$

Selecting maximal empirical AUROC loses at most 2ϵ relative to the best candidate for that bank. An observed gap greater than 2ϵ certifies the same population ordering at this confidence level.

The trusted bank and evaluation labels both count towards annotation cost. This guarantee concerns the supplied score functions for that bank; it does not bound the error caused by using a sparse bank in place of the full field. Our finite-dataset resampling audit reports selection frequencies rather than claiming the independent-sample guarantee.

J.2 From source reconstruction to detector decisions

Exact source sufficiency specifies which evidence can be reconstructed. To examine how reconstruction affects detector decisions, we use the four cached disjoint CIFAR-100 bundles. Before computing the new outcomes, we fix dimensions $m \in \{1, 2, 8, 32, 64, 99\}$ and three nested encoders: leading singular directions of the distinct-contrast matrix (obtained by singular value decomposition, SVD), a centred random orthonormal basis with one fixed seed, and a fixed Helmert contrast basis. The first minimises contrast-matrix Frobenius error. All three preserve every class contrast at 99 channels.

Features, recorded labels, $k = 100$ neighbours, fitted heads, and baseline scores stay fixed. We propagate encoded one-hot labels and decode their neighbour mean, restoring the known constant $1/C$. Queries either compare the recorded label with its fixed head-selected rival or take the strongest alternative, as in the original label-conflict detector. Compressed fields are not clipped, renormalised, or recalibrated. To preserve exact numerical controls, decoded values within 10^{-12} of a neighbour-count grid value are snapped to that value before the original float32 readout.

Bounds on decision changes. Let F_i and $\widehat{F}_i$ be the original and decoded label fields, $\delta_i = \|\widehat{F}_i - F_i\|_\infty$, and $\epsilon_i = 2\delta_i$. For either query, its score error satisfies

$$|\widehat{s}_i - s_i| \le \epsilon_i.$$

For a fixed contrast, apply the triangle inequality to its two entries; for the maximum query, use that a maximum is 1-Lipschitz in the sup norm. Consequently, signs are preserved whenever $|s_i| > \epsilon_i$, and a pair's ordering is preserved whenever $|s_i - s_j| > \epsilon_i + \epsilon_j$. For corrupted indices P and clean indices N, this also gives

$$|\widehat{\mathrm{AUC}}(\widehat{s}) - \widehat{\mathrm{AUC}}(s)| \le \frac{1}{|P||N|} \sum_{i \in P, j \in N} \mathbf{1}\{|s_i - s_j| \le \epsilon_i + \epsilon_j\}.$$

Only pairs satisfying this inequality can change their AUROC contribution, by at most one each. A top-budget review set is unchanged if the ordering condition holds for every selected–unselected pair. These sufficient certificates depend on actual field errors and decision margins. Matrix reconstruction error or source dimension alone does not supply the quantities needed to evaluate them.

Channels	Pair: SVD	Pair: random	Max: SVD	Max: random	Max: Helmert
1	−30.84	−23.55	−33.25	−32.75	−33.20
2	−30.05	−20.11	−32.02	−28.64	−33.48
8	−24.00	−9.36	−27.77	−10.14	−31.28
32	−15.82	−2.56	−16.53	−1.81	−20.94
64	−7.48	−0.71	−7.51	−0.37	−10.63
99	0.00	0.00	0.00	0.00	0.00

Table 8: Compression-induced AUROC changes in pp, relative to the corresponding full query, averaged equally over four cached groups. Pair is the fixed-rival query; Max is the original strongest-alternative query. All dimensions and encoders were fixed before these outcomes.

Reconstruction error and detection performance. At 64 channels, SVD has the lower contrast reconstruction error (51.28% versus 59.59% for the random encoder), but loses much more max-query AUROC: 7.51 pp versus 0.37 pp. At a 10% review budget, SVD replaces 36.28% of selected examples and reduces precision by 9.85 pp on average; the random encoder replaces 15.05% and

changes precision by -0.15 pp. The Helmert control loses 10.63 pp AUROC and 13.00 pp review precision. Under unchanged five-bin, weight-two fusion with every stored baseline, mean max-query AUROC changes are respectively -2.35, -0.11, and -3.12 pp. These are changes relative to the same baseline fused with the full query, not relative to the baseline alone.

Group	Full AUROC	SVD Δ	Random Δ	SVD review Δ	Random review Δ
Human, partition 0	78.30	-6.66	-0.23	-11.70	$+1.90$
Human, partition 1	78.91	-7.22	-0.11	-11.00	$+0.80$
Synthetic, partition 0	91.43	-7.92	-0.65	-7.90	-2.10
Synthetic, partition 1	90.97	-8.26	-0.47	-8.80	-1.20

Table 9: Group-level max-query results at 64 channels. Full AUROC is a percentage; changes are pp. Review changes are precision changes at the fixed 10% budget. Shared partitions are not independent datasets.

Decision changes by margin. At 64 channels, SVD changes 16.43% of max-query signs and strictly reverses 18.96% of the fixed sampled score pairs. With the random encoder, no sign changes occur among the 16,552 queries whose original score magnitude is at least 10 pp; SVD changes 2,008 of these signs (12.13%). Both encoders break all 1,136 zero-score ties, illustrating why tie changes must be reported separately from strict reversals. Measured field-error certificates cover 25.06% of signs and 10.86% of sampled rankings for SVD; for the random encoder, coverage is 42.15% and 14.79%. No changes occur outside the vulnerable sets, using the declared numerical tolerances. Coverage is conservative; the study does not evaluate the full clean–corrupted-pair AUROC bound above.

Interpretation and provenance. The 99-channel controls reproduce both full score arrays exactly for every encoder and group. Thus lower-dimensional differences are not artefacts of the full-control readout. The study shows that compression can affect actual detector decisions, but also that exact source sufficiency is stronger than near-preservation of detection quality. It does not rule out effective approximate compression below 99 channels. The random encoder uses one fixed seed; its result is not a general superiority claim.

The protocol and input hashes precede score generation, and score hashes precede corruption evaluation. This remains a retrospective study on previously evaluated data, outside the historical prediction count. The driver `compression_decision_audit.py` retains all 144 query outcomes, review budgets 5%, 10%, and 20%, baseline-specific fusion changes, margin strata, and 200,000 fixed nonself ranking pairs per group. Independent rank-sum AUROC and stable-sort review checks in `validate_compression_decision.py` agree with all 144 outcomes and 432 review comparisons; all 24 exact controls pass. No efficiency gain or fresh dataset replication is claimed.

Random-encoder sensitivity. A separate retrospective extension fixes ten seeds, 71201–71210, including the original 71201, before generating its new scores. All four groups, neighbours, heads, and queries stay unchanged; only the random encoder varies. At 64 channels, the strongest-alternative query loses 0.63 pp AUROC on average, with four-group seed means ranging from 0.36 to 0.92 pp loss. Review precision at a 10% budget loses 0.94 pp on average, with seed means from 0.15 to 2.08 pp loss. The nine additional seeds alone lose 0.66 pp AUROC and 1.03 pp review precision. The original encoder was favourable, but the contrast with SVD persists: each random encoder has higher AUROC than SVD in each group (40 dependent group–seed comparisons). These are encoder repetitions on four reused groups, not new datasets.

Mean contrast reconstruction error is 59.48%, versus SVD's 51.28%. All 80 full-dimensional query controls reproduce the full scores exactly, and the original seed's 64-channel scores are reproduced exactly. The separate protocol `REVIEW_ROBUSTNESS_PROTOCOL.md` and driver `review_robustness_audit.py` preserve the historical audit unchanged. Scores are exported and hashed before evaluation. Independent rank-sum and selection checks in `validate_review_robustness.py` reproduce all 160 compression outcomes and 135 gradient-subset evaluations, with maximum AUROC discrepancy below 3×10^{-14} pp.

J.3 SOURCE SUFFICIENCY AND READOUT CORRECTNESS

The source characterisation determines whether the sources retain the distinctions needed by a query. The compression audit additionally examines whether the chosen readout uses that information correctly. Under an inner-product readout, a class source with a small norm can give another class more support than itself, producing a source–query mismatch even at the population level.

A three-class readout example. Encode bird, dog and cat as $r_b = (0.1, 0)$, $r_d = (0.2, 1)$ and $r_c = (-0.3, -1)$. In a pure bird field $w = r_b$, the inner-product readout gives bird 0.01 and dog 0.02, flagging a correct bird label. These sources sum to zero. Their encoder has rank two and its kernel is the constant direction, so every three-class contrast is recoverable. The third class permits unequal source lengths while retaining centring. The incorrect bird vote therefore results from the chosen readout, despite source sufficiency.

The diagnostic and repair. Let $R = [r_1 \ \cdots \ r_C]$ satisfy $R\mathbf{1} = 0$. In a homogeneous class-t context, suppose a recorded label stays correct with probability $1 - \eta$ and is otherwise uniform over the other classes. For $\eta < (C-1)/C$, the population encoded field is

$$w_t = \beta r_t, \qquad \beta = 1 - \frac{C\eta}{C-1} > 0.$$

The raw readout assigns class c the vote $r_c^\top w$. Its noise score is the strongest alternative vote minus the recorded label's vote.

Proposition 7 (Self-support and the stability of a directional query)**.** *Write $P = R^\top R$. At $w_t = \beta r_t$, the raw readout uniquely selects class t exactly when*

$$P_{tt} > \max_{c \neq t} P_{ct}.$$

Assume all sources are nonzero and their unit directions $d_c = r_c/\|r_c\|$ are distinct. Replacing the class vote by $d_c^\top w$ uniquely selects t at every w_t, for every $\beta > 0$. Its exact distance from w_t to the nearest class-decision boundary is

$$\rho_t = \beta \|r_t\| \min_{c \neq t} \sqrt{\frac{1 - d_t^\top d_c}{2}}.$$

Every perturbation e with $\|e\| < \rho_t$ preserves that decision and the correct clean/noisy sign for every recorded label.

Proof. The raw votes at w_t are βP_{ct}, giving the first condition. The directional true-class vote is $\beta\|r_t\|$, whereas rival c receives $\beta\|r_t\| d_c^\top d_t$, a strictly smaller value. The gap to c divided by the boundary normal length $\|d_t - d_c\| = \sqrt{2(1 - d_t^\top d_c)}$ gives its hyperplane distance. Minimising over rivals gives the largest ball contained in the true-class decision region. Moving along a closest negative normal reaches a tie at ρ_t and defeats the true class beyond it. When the true class wins strictly, the recorded label has a negative conflict score exactly when it is the true class. □

Under these assumptions, normalising the readout removes homogeneous self-support failures. The radius quantifies the remaining sensitivity to small source lengths and nearly aligned class directions. Together with source sufficiency, this distinguishes information loss from incorrect readout and from a correct decision with limited tolerance to perturbation.

A fixed repair of the compression failure. The four 64-channel SVD encoders fail the raw self-support condition for 14, 6, 12, and 10 classes, in the group order below. The original random encoder fails for none. Query-direction normalisation removes every such population failure while keeping sources, neighbour graphs, recorded labels, and the original fitted heads unchanged. The repair requires no clean labels, noise-rate estimate, new prediction head, or fitted mixing weight.

Mean SVD AUROC rises from 77.39% to 79.26%; the repaired random control attains 84.66% and the full label field 84.90%. Mean SVD review precision improves by 0.83 pp, with one group losing 0.10 pp. Thus the homogeneous self-support result supplies a specific repair whose benefit transfers partially to heterogeneous fields. The remaining SVD deficit is consistent with other compression effects; the radius alone does not attribute or predict its magnitude.

Group	SVD AUROC gain	Random AUROC gain	SVD review gain
Human, partition 0	+2.10	+0.17	+2.50
Synthetic, partition 0	+1.98	+0.16	+0.60
Human, partition 1	+1.65	+0.07	+0.30
Synthetic, partition 1	+1.75	+0.10	−0.10

Table 10: Query-only repair on the cached compression groups. Changes are percentage points relative to each encoder's raw readout; review uses a 10% budget. The random control uses the original fixed seed. Groups are reused development data, with shared partitions.

Scope and provenance. Singh-Miller & Collins (2009) already average embedded neighbour labels and score classes by inner products with label prototypes, learning the embeddings through leave-one-out conditional likelihood. Our compression audit instead fixes the encoder and graph, then asks whether the readout can select the wrong class from sufficient sources. Proposition 7 answers through an exact self-support condition and the geometry of the query-only repair. Normalised directions and hyperplane-distance margins are standard constructions; their role here is to distinguish information loss from a correctable readout failure and quantify the repaired readout's perturbation tolerance. The matched intervention tests this distinction on cached detector fields. Its heterogeneous-data AUROC gain is measured separately from the homogeneous-field guarantee.

The driver `source_query_geometry.py` fixes the intervention and hashes inputs before generating its new scores; score bundles are hashed before corruption evaluation. All eight raw arrays reproduce their earlier compression scores exactly. Independent rank-sum and review-selection checks agree with the reported metrics. These are retrospective mechanism and repair results on four previously evaluated groups, outside the historical prediction count. The theorem concerns homogeneous fields; the experiment measures transfer of its repair to sampled heterogeneous neighbourhoods. It establishes neither a source-selection rule nor a measured reduction in experimentation cost.

J.4 Risk and ranking consequences

Theorem 7 (Plug-in detection bound)**.** *Let $q^\star(U) = \mathbb{P}(D = 1 \mid U)$, let $g^\star = \mathbf{1}\{q^\star \geq 1/2\}$, and let $\widehat{g} = \mathbf{1}\{\widehat{q} \geq 1/2\}$. Then*

$$R(\widehat{g}) - R(g^\star) \leq 2\mathbb{E}|\widehat{q} - q^\star|. \tag{18}$$

For ranking, let $\pi = \mathbb{P}(D = 1) \in (0, 1)$ and let U, U' be independent. Define $L_s(U, U')$ to be one when s orders a pair opposite to $q^\star$, one half when s ties a pair with unequal posteriors, and zero otherwise.

Proposition 8 (AUROC regret)**.** *With ties awarded one half, the regret relative to the Bayes ranking is*

$$\mathrm{Reg}_{\mathrm{AUC}}(s) = \frac{1}{2\pi(1-\pi)}\mathbb{E}[|q^\star(U) - q^\star(U')|L_s(U, U')]\,. \tag{19}$$

If $\|s - q^\star\|_\infty \leq \epsilon$, only pairs with posterior gap at most 2ϵ can contribute, and therefore

$$\mathrm{Reg}_{\mathrm{AUC}}(s) \leq \frac{\mathbb{E}[|q^\star - q^{\star\prime}|\mathbf{1}\{|q^\star - q^{\star\prime}| \leq 2\epsilon\}]}{2\pi(1-\pi)}. \tag{20}$$

The bound relates ranking performance to the location of approximation errors: regret arises from reversed or tied pairs with unequal posteriors, with larger posterior gaps contributing more.

J.5 Component error decomposition

Let $q^{(0)} = q_X$ be the full-input posterior; $q^{(1)} = q_Z$ the representation-limited posterior; $q^{(2)}$ the posterior under the identifiable noise model; $q^{(3)}$ the chosen population field; $q^{(4)}$ the finite-data

query; and $q^{(5)} = \widehat{q}$ the calibrated output. The triangle inequality gives

$$\begin{aligned}\mathbb{E}|\widehat{q} - q_X| \le \underbrace{\mathbb{E}|q_X - q_Z|}_{\Delta_{\text{repr}}} + \underbrace{\mathbb{E}|q_Z - q^{(2)}|}_{\Delta_{\text{ident}}} + \underbrace{\mathbb{E}|q^{(2)} - q^{(3)}|}_{\Delta_{\text{field}}} \\ + \underbrace{\mathbb{E}|q^{(3)} - q^{(4)}|}_{\Delta_{\text{query}}} + \underbrace{\mathbb{E}|q^{(4)} - \widehat{q}|}_{\Delta_{\text{cal}}}.\end{aligned} \tag{21}$$

Equation (21) separates failure modes requiring different interventions. Transition identification requires assumptions or additional observations, field estimation depends on the propagation rule, and query bias and calibration require changes to the readout.

For a concrete rate, suppose T is known, constant, and nonsingular; the observed-label field is α-Hölder in a d-dimensional representation; the density is locally bounded below; and a regular bounded kernel of bandwidth h is used. If the posterior transformation is bounded away from zero, the oracle-transition kernel query satisfies the standard form

$$\mathbb{E}|q^{(4)} - q_Z| = O\Big(\big\|T^{-1}\big\| \left[h^\alpha + (nh^d)^{-1/2}\right]\Big). \tag{22}$$

The two terms are field-smoothing bias and finite-neighbourhood variance. Their interaction with Theorem 7 and Proposition 8 converts query convergence into detection guarantees.

K Neural Training as an Exact Dynamic Field

Let $f_\theta(x_i) \in \mathbb{R}^C$ be the logits of any differentiable network, $J_i = \partial f_\theta(x_i)/\partial\theta \in \mathbb{R}^{C\times P}$ its parameter Jacobian, and $p_i = \mathrm{softmax}(f_\theta(x_i))$. Define the matrix-valued empirical tangent block

$$\Theta_{ij}(\theta) = J_i(\theta)J_j(\theta)^\top \in \mathbb{R}^{C\times C}. \tag{23}$$

Unlike an infinite-width neural tangent kernel analysis (Jacot et al., 2018), we allow this operator to evolve with the finite network.

Theorem 8 (Exact training-field identity)**.** *Under gradient flow on mean softmax cross-entropy with observed labels,*

$$\dot\theta = -\frac{1}{n}\sum_{j=1}^{n} J_j^\top (p_j - e_{\widetilde{y}_j}), \tag{24}$$

the logit velocity at every example is

$$\dot f_\theta(x_i) = -\frac{1}{n}\sum_{j=1}^{n} \Theta_{ij}(\theta)(p_j - e_{\widetilde{y}_j}). \tag{25}$$

Thus $p_j - e_{\widetilde{y}_j}$ is an example-local label source and Θ_{ij} is the exact learned propagator from j to i at the current state.

Proof. The chain rule gives $\dot f_\theta(x_i) = J_i\dot\theta$. Substitute the parameter velocity and use $J_iJ_j^\top = \Theta_{ij}$. □

Corollary 4 (Label replacement response)**.** *At a common network state, replacing only example j's label y_j by c changes the initial logit velocity at i by*

$$\dot f_i^{(c)} - \dot f_i^{(y_j)} = \frac{1}{n}\Theta_{ij}(e_c - e_{y_j}). \tag{26}$$

The common probability vector cancels in this difference, so the label-replacement response is determined by the tangent block applied to the label contrast.

Integrating Equation (25) over time yields the prediction trajectory. When Θ_t changes, the field coordinates and propagator co-evolve; when it is approximately fixed, familiar kernel dynamics result. Either way, the instantaneous identity remains exact almost everywhere along a differentiable training path.

The operator separates the example's own contribution from contributions by other examples:

$$\dot{f}_i = -\frac{1}{n}\Theta_{ii}(p_i - e_{\widetilde{y}_i}) - \frac{1}{n}\sum_{j\neq i}\Theta_{ij}(p_j - e_{\widetilde{y}_j}). \tag{27}$$

The first term is the example's self-field, and the second is its contextual field. A corrupted example can eventually be memorised through its self-field even when its observed label conflicts with the contextual response. This decomposition provides an operator interpretation of the early-learning and memorisation phenomena used by dynamics-based detectors.

Proposition 9 (Margin detectors are temporal field queries)**.** *Let*

$$m_i(t) = f_{\widetilde{y}_i}(x_i, t) - \max_{c\neq\widetilde{y}_i} f_c(x_i, t)$$

and suppose the maximising competitor $c_i(t)$ is unique almost everywhere. Set $a_i(t) = e_{\widetilde{y}_i} - e_{c_i(t)}$ and $r_j(t) = p_j(t) - e_{\widetilde{y}_j}$. Then

$$\dot{m}_i(t) = -\frac{1}{n}\sum_j a_i(t)^\top \Theta_{ij}(t) r_j(t) \quad \textit{almost everywhere}. \tag{28}$$

Consequently, the continuous-time area under the margin (Pleiss et al., 2020) over $[0, T]$ is

$$\frac{1}{T}\int_0^T m_i(t)\,dt = m_i(0) - \frac{1}{nT}\int_0^T (T-t)\sum_j a_i(t)^\top \Theta_{ij}(t) r_j(t)\,dt. \tag{29}$$

Forgetting events (Toneva et al., 2019) are downward zero crossings of the same queried margin path.

Proof. Where the maximising competitor is unique, differentiate the margin and substitute Equation (25). The set of tie times has measure zero by assumption. Integrating $m_i(t) = m_i(0) + \int_0^t \dot{m}_i(s)ds$ over $t \in [0, T]$ and exchanging the order of integration gives Equation (29). □

Because Θ_{ij} can act differently on different label directions, a scalar trace, Frobenius norm, or feature-space cosine similarity generally does not determine the label-replacement response in Equation (26). Consequently, geometrically close examples can have different effects on a queried label.

For a linear classifier block $f_i = Wh_i + b$, the part of the tangent operator contributed by W is

$$\Theta_{ij}^W = (h_i^\top h_j) I_C. \tag{30}$$

Consequently, its contextual velocity is a similarity-weighted mean of the full class-residual vectors. This restricted operator motivates the query in Section L and identifies the directional information omitted by a norm-only residual source.

L Scores Used in the Polarity Experiment

The polarity experiment compares direct label evidence, residual work and a fixed combination of the two. These scores share frozen coordinates, recorded labels and out-of-sample predictions; changing the residual readout isolates the effect of the orientation repair.

Let z_i be a frozen pretrained coordinate, p_i a two-fold out-of-sample predictive distribution, and $r_i = p_i - e_{\widetilde{y}_i}$. For the $k = 100$ cosine-nearest neighbours $N_k(i)$, excluding i, define

$$\alpha_i = \frac{1}{k}\sum_{j\in N_k(i)} e_{\widetilde{y}_j}, \qquad s_i^{\mathrm{lab}} = \max_{c\neq\widetilde{y}_i}\alpha_{i,c} - \alpha_{i,\widetilde{y}_i}.$$

Label conflict asks whether neighbouring recorded labels support an alternative class. For a response bank B_i, residual work is

$$\bar{r}_i^B = \frac{1}{|B_i|}\sum_{j\in B_i} r_j, \qquad s_i^{\mathrm{work}} = -r_i^\top \bar{r}_i^B.$$

It measures opposition between the example's residual and the contextual residual field, as motivated by Equation (30). The pilot-based repair replaces this work score using Equation (32).

For $n > 1$ scored examples, define the average-rank transform

$$R(v_i) = \frac{\mathrm{rank}_{\mathrm{average}}(v_i) - 1}{n - 1}.$$

The combined score used in the experiment is

$$s_i^{\mathrm{combined}} = R(s_i^{\mathrm{lab}}) + 0.25R(s_i^{\mathrm{work}}),$$

with ranks computed across the scoring population and ties retained. The coefficient is fixed before the polarity test. The reference version uses all 100 neighbours for work; the bank-matched raw and repaired versions use the same 50-example response bank and retain the 100-neighbour label term. Section M.2 specifies the bank split and reports every predeclared criterion, including whether the combined score improves and whether direct label evidence remains stronger.

M Confidence Reversal: Mechanism, Proofs and Tests

Consider bird images with independent 40% label flips. Predictions assigning 55% probability to bird give a noisy-minus-clean contextual-work gap of 0.1; sharpening to 80% gives a gap of -0.4. Both models classify the images as birds. The observed bird frequency is 60%, so the sharpened model's mean residual corresponds to an update that reduces bird confidence. The clean bird's gradient opposes this update, so the opposition score now assigns greater suspicion to the correct label.

Theorem 9 locates the reversal exactly in the homogeneous binary model. It also predicts that increasing the reference-bank size strengthens the resulting orientation. A disjoint pilot bank supplies the targeted repair in Corollary 5. The source–query distinction yields a second prediction: normalising each binary gradient before aggregation removes positive temperature rescalings, whereas normalising their mean retains the confidence-dependent weights (Proposition 10).

Figure 2 displays the exact prediction and both empirical tests.

The cross-method test and reference-bank audit separate confidence weighting from reference composition. Pairwise cosine remains temperature-invariant; replacing mixed-class banks by matched same-true-class banks isolates the additional effect of propagation geometry. These results, including the oracle-bank construction and all comparison outcomes, are retained in Sections M.3 and M.6.

M.1 Confidence-dependent orientation of contextual work

A residual measures the difference between predictions and labels. Its population field is $\bar{p} - \alpha$. Changing prediction confidence can therefore reverse this field while leaving the label geometry unchanged. For the contextual-work query, this implies that more accurate averaging can worsen detection when the query has the incorrect orientation.

Theorem 9 (Contextual-work polarity and finite-sample separation)**.** *Consider a homogeneous binary neighbourhood with independent label flips at rate $\eta \in (0, 1/2)$. Predictions assign probability $(1 + \lambda)/2$ to the true class, with fixed $\lambda \in [0, 1]$. Let $a = 1 - 2\eta$ and compute contextual work from k reference labels independent of the query's flip. For AUROC draws with independent reference banks, let U, V be independent* $\mathrm{Binomial}(k, 1 - \eta)$ *variables. Then*

$$A_k(\lambda) = \mathbb{E}H((1 + \lambda)(2U/k - 1) + (1 - \lambda)(2V/k - 1) - 2\lambda), \tag{31}$$

where H awards half credit to ties. Writing $b_k = \exp\{-k(a - \lambda)^2/(1 + \lambda^2)\}$,

$$\lambda < a \implies A_k \geq 1 - b_k, \qquad \lambda > a \implies A_k \leq b_k.$$

At $\lambda = a$, $A_k \to 1/2$ as $k \to \infty$.

Thus confidence below the observed-label contrast gives the correct polarity; confidence above it gives the opposite polarity. Increasing k sharpens either conclusion. Every $\lambda > 0$ has the same classification decision in this model, so the intervention isolates confidence at fixed classification accuracy. Above the threshold, the fixed opposition query assigns higher population scores to clean examples, despite the field retaining label information. The theorem applies specifically to this contextual residual-opposition query.

Normalisation before and after aggregation. For binary cross-entropy at a fixed network state, temperature rescales each example's parameter gradient by a positive scalar. Pairwise gradient cosines are therefore unchanged, whereas the direction of their unnormalised sum can change (Proposition 10). This distinguishes averaging cosine similarities from taking a cosine after averaging gradients. AGRA's cross-entropy (CE) comparison uses the latter operation (Sedova et al., 2023); gradient-cosine controls use the former (Nguyen-Duc et al., 2023). This analysis characterises the comparison signal; performance under adaptive training also depends on the resulting updates and selection feedback.

An observable polarity repair. Use a separate pilot bank to form $\widehat{\alpha}$ and $\widehat{\overline{p}}$. With uniform class vector u, define

$$h_i = \operatorname{sign}\left[(\widehat{\alpha}_i - u)^\top (\widehat{\alpha}_i - \widehat{\overline{p}}_i)\right], \qquad s_i^{\mathrm{pol}} = h_i s_i^{\mathrm{work}}, \tag{32}$$

using $\operatorname{sign}(0) = 0$. This requires no clean labels. In the homogeneous model, h_i estimates $\operatorname{sign}(a-\lambda)$.

Corollary 5 (Finite-pilot repair)**.** *Under Theorem 9, use an independent m-reference pilot for each query in Equation* (32)*. For $d = |a - \lambda| > 0$, the corrected query satisfies*

$$A_{k,m}^{\mathrm{pol}} \geq 1 - b_k - 2e^{-md^2/2} - 2e^{-ma^2/2}.$$

The last term accounts for choosing the wrong class orientation from the pilot. No positive separation is guaranteed at $d = 0$: calibration removes the population residual field. The proof and exact-binomial checks appear in Section M.7. The label field already separates the homogeneous population, so the repair corrects the query's use of existing information.

M.2 STL-10 PROTOCOL AND OUTCOMES

Before scoring, the first experiment fixes all 8,000 STL-10 test images (Coates et al., 2011), with class IDs decremented to 0–9 and pairs $(0, 1), (2, 3), \ldots, (8, 9)$. For pair index p and repetition $r \in \{0, 1, 2\}$, independent Bernoulli 40% label flips use seed $19300 + 100p + r$; two-fold noisy-label multilayer perceptron (MLP) heads use $29300 + 100p + r$. Frozen ResNet-18 features and the head architecture are specified in Section R. Heads train for ten epochs with learning rate 0.05, momentum 0.9 and batch size 256. Applying temperatures $\{0.25, 0.5, 1, 2, 4\}$ to binary log odds after fitting preserves predicted classes and confidence rankings.

Alternating entries of each ordered 100-neighbour list form disjoint 50-example pilot and response banks. The repair uses Equation (32), with sign zero defined as zero. The reference combined score uses 100-neighbour work; the bank-matched raw control and repair use the same 50-reference response bank. All retain the 100-neighbour label term and residual weight 0.25 (Section L). The frozen criteria and their outcomes are collected once in Table 15.

Raw-work AUROC falls from 90.22% at $T = 4$ to 24.18% at $T = 0.25$; repair restores 85.73% (Figure 2). The combined score rises from 95.67% to 98.67%, while confidence remains 80.96%. Against the bank-matched raw combined score, repair gains 2.795 pp at $T = 0.25$.

Direct label conflict reaches 99.10% and exceeds the repaired combined score in every sharp-temperature run. At $T = 4$, repair reduces work AUROC in all 15 runs and the combined score in 13/15, with mean full-score change -0.137 pp. No temperature is selected for deployment. The construction uses clean labels only to generate corruption and evaluate scores. Heterogeneous neighbourhoods and learned probabilities relax the theorem's homogeneity and independence assumptions: disjoint bank indices alone do not imply statistical independence. STL-10's ImageNet origin also permits encoder pretraining overlap. Seeds and temperatures share images.

The separate cross-method test in Section M.6 applies the confidence intervention to AGRA's comparison signal, with fresh corruptions and ordinary linear-head training.

Ordinary checkpoints: a within-example diagnostic. We additionally audit all 60 saved binary linear-head checkpoints at $T = 1$, without sharpening or selecting epochs. Let j_i be the binary logit's parameter Jacobian, $\widehat{j}_i = j_i/\|j_i\|$, and F_i the mean gradient of the fixed self-excluding reference bank. Holding the model and bank fixed, the AGRA-CE comparison's noisy-minus-clean score gap

at the same example is

$$\Delta_i = -2(2y_i^\star - 1)\widehat{j}_i^\top \frac{F_i}{\|F_i\|}.$$

Use zero when the field is zero. To verify the identity, the clean-label gradient direction is $-(2y_i^\star - 1)\widehat{j}_i$ and flipping the binary label reverses it. Substituting both directions in the negative cosine gives the formula. Thus $\Delta_i < 0$ means the score prefers the flipped label as cleaner at this image. Clean labels are diagnostic inputs only.

At epochs 1, 5, 10, and 30, this occurs for respectively 49.55%, 49.68%, 49.83%, and 49.79% of queries, averaged over the 15 runs. The corresponding corruption AUROCs are 58.48%, 50.42%, 49.97%, and 49.99%. Replacing the raw reference field by the mean of unit gradients still gives reversed within-example ordering on 45.35% of queries at every epoch. The reconstruction matches saved AGRA-CE scores within 4.45×10^{-16}.

Epoch	Reversed ordering: median [IQR]; range	AUROC: median [IQR]; range
1	49.75 [48.91, 50.00]; 47.81–51.63	56.22 [54.23, 60.73]; 48.85–72.41
5	50.00 [49.16, 50.50]; 47.56–51.44	50.47 [49.09, 53.50]; 32.91–60.75
10	50.00 [49.34, 50.25]; 47.19–51.50	50.34 [48.93, 53.30]; 33.36–60.50
30	49.94 [49.31, 50.19]; 48.25–52.19	50.85 [47.79, 53.11]; 33.10–60.33

Table 11: Distribution over all 15 runs per ordinary checkpoint, in percent. IQR denotes the 25th–75th percentile interval. Shared images and repeated checkpoints do not supply independent replications.

For the unit-gradient control, the reversal median is 45.06%, IQR $[44.00, 46.31]\%$, and range 42.44%–49.69% at every epoch. Table 11 and the per-run results report the variation around these means, especially in corruption AUROC.

Within-example counterfactual ordering and cross-example AUROC measure different properties; the former does not determine the latter. The matched intervention below separates reference composition from confidence-dependent weights in these fixed-state binary comparisons.

M.3 Separating reference composition from confidence

We reuse all 15 saved STL-10 runs and checkpoints 1, 5, 10, and 30. For every query, replace its original 256-member mixed-class bank with 256 distinct examples of the query's true class, excluding the query. Match the original bank's number of corrupted labels exactly, sampling without replacement with seed $91300 + 100\,\mathrm{pair} + \mathrm{seed}$. The same banks serve every checkpoint and temperature $T \in \{0.25, 1, 4\}$. Cross this membership intervention with native gradient magnitudes and unit source magnitudes. The unit-source score is pairwise cosine; further positive normalisation of its aggregate preserves within-example orientation, but can change cross-example rankings. Clean labels construct oracle diagnostic banks; the resulting scores are not deployable detector estimates.

With unit sources, changing bank composition reduces misorientation in every class pair. The overall rate falls from 45.35% to 0.08%; the five pair means fall from a range of 43.65–48.27% to 0–0.23%. Unit-source AUROC rises from 56.68% to 99.999%. These outcomes are identical across checkpoints and temperatures: the fixed binary linear-head Jacobians and recorded-label signs determine the unit directions. This confirms the audit's recorded empirical hypothesis that class composition contributes to the invariant control's misorientation.

	Misorientation (%)		AUROC (%)	
Epoch	Mixed	Same-class	Mixed	Same-class
1	49.55	33.42	58.48	78.46
5	49.68	41.47	50.42	63.33
10	49.83	44.48	49.97	58.72
30	49.79	45.50	49.99	57.86

Table 12: Matched reference composition with native gradient magnitudes at $T = 1$. Means over the same 15 runs; repeated checkpoints share images.

With native gradient magnitudes, substantial misorientation remains (Table 12). At epoch 30, sharpening from $T = 4$ to $T = 0.25$ increases same-class misorientation from 0.92% to 83.52% and reduces AUROC from 99.93% to 11.44%. Counterfactual orientation changes for 82.60% of queries, versus 27.47% under mixed banks. Thus homogeneous class composition can amplify the temperature response even while fixing most unit-source misorientation. The intervention controls corruption count, but changes both Jacobian directions and label signs. Its unit-source arm removes confidence-dependent weights; it does not isolate feature geometry alone. Same-class images retain heterogeneous predictions, beyond the theorem's constant-confidence model.

This is a retrospective mechanism audit with one new bank draw per saved run. Protocol, driver, and input SHA-256 digests were recorded before new comparisons; exported score bundles were hashed before metric evaluation. These hashes identify content rather than independent timestamps. The driver `reference_geometry_audit.py` exports all 720 metric rows and paired temperature contrasts. Reconstructed original scores and independent scalar checks agree within 2.86×10^{-15}. Independent tests verify the counterfactual signs, bank constraints, digests, and rank-sum AUROCs.

M.4 Sensitivity to near-zero gradient aggregates

The cosine of an exactly zero aggregate is undefined; our implementation assigns it zero. To distinguish vector cancellation from the overall temperature-dependent gradient scale, define

$$q_i(T) = \frac{\| \operatorname{mean}_{j \in B_i} g_j(T) \|}{\operatorname{mean}_{j \in B_i} \| g_j(T) \|}.$$

This ratio lies in $[0, 1]$ for nonzero reference magnitudes. A separate retrospective audit reconstructs all 15 cross-method runs at epoch 30 and all five stored temperatures. No aggregate is exactly zero; the smallest $q_i(T)$ is 1.95%. Every reconstructed AGRA score agrees with its saved value to 10^{-10}.

For each run, we retain the same examples at both endpoint temperatures: those whose ratio exceeds the threshold at both $T = 4$ and $T = 0.25$. We also remove fixed fractions of the smallest endpoint-minimum ratios. Table 13 reports every specified filter. All retained subsets contain both clean and corrupted examples. Pairwise-cosine control scores remain invariant within 1.12×10^{-16}.

Endpoint filter	Retained (%)	Mean gap (pp)	Minimum run gap (pp)
$q > 0$	100.00	27.69	11.51
$q > 0.001$	100.00	27.69	11.51
$q > 0.01$	100.00	27.69	11.51
$q > 0.05$	60.88	31.53	17.33
$q > 0.10$	13.89	38.90	27.83
Remove lowest 1%	99.00	27.83	11.59
Remove lowest 5%	95.00	28.08	12.18
Remove lowest 10%	90.00	28.46	12.09
Remove lowest 20%	80.00	29.42	12.88

Table 13: Cancellation sensitivity on identical retained examples at both temperatures. Gap is AGRA-comparison AUROC at $T = 4$ minus that at $T = 0.25$; means weight the 15 runs equally. Ratio thresholds are dimensionless.

The temperature response persists after excluding the most cancellation-prone aggregates. This check evaluates the sampled endpoint scores; behaviour between those endpoints would require analysis of the continuous temperature path. Filtered AUROCs describe the selected populations and provide retrospective sensitivity checks. The audit does not attribute ordinary-checkpoint failures uniquely to confidence. Full norms, ratios, retention, class counts, and per-run outcomes are retained by `review_robustness_audit.py` under the separate robustness protocol.

M.5 Normalisation order and source confidence

Proposition 10 (Binary gradient direction invariance)**.** *Fix a differentiable binary classifier with logit difference $f_\theta(x)$, observed labels $y_i \in \{0, 1\}$, and finite logits. For $T > 0$, let $g_i(T)$ be the parameter gradient of binary cross-entropy applied to $f_\theta(x_i)/T$. If $\nabla_\theta f_\theta(x_i) \neq 0$, then $g_i(T) = c_i(T) g_i(1)$*

for $c_i(T) > 0$. Consequently every fixed aggregation of pairwise gradient cosines is temperature invariant. No such conclusion follows for $\cos(g_i(T), \sum_j w_{ij} g_j(T))$, even with fixed nonnegative weights.

Indeed,

$$g_i(T) = \frac{\sigma(f_\theta(x_i)/T) - y_i}{T} \nabla_\theta f_\theta(x_i).$$

The scalar has a fixed nonzero sign determined by y_i. Positive scaling cancels in each pairwise cosine, but changes the relative source weights in the unnormalised sum. For a counterexample, take a homogeneous binary neighbourhood with a common nonzero logit Jacobian, true label one, and noisy observed-label fraction $\eta < 1/2$. Its population gradient is proportional to $\lambda - (1 - 2\eta)$. On crossing $\lambda = 1 - 2\eta$, the sign of the sample-to-field cosine reverses for both clean and noisy queries. At the crossing the population comparison vector is zero and its cosine is undefined; assigning zero is an implementation convention at this unseparated population state. Normalising sources before aggregation avoids this scalar reweighting, but can discard useful magnitude information. Neither better detection nor invariance under retraining or multiclass temperature scaling follows.

For a linear logit on fixed features, its Jacobian is parameter independent. Pairwise cosine then depends only on observed label signs and feature directions, making the score independent of the learned head parameters. The resulting invariance follows from omitting magnitude information; its detection performance depends on the remaining directional evidence.

M.6 Cross-method polarity confirmation

After the first STL-10 experiment, we freeze corruption seeds $49300 + 100p + r$, model seeds $59300 + 100p + r$, and comparison-bank seeds $69300 + 100p + r$, using its five pairs and three repetitions. Binary linear heads fit ordinary noisy-label cross-entropy on row-normalised frozen features. Stochastic gradient descent (SGD) uses learning rate one, momentum 0.9, batch size 128, and 30 epochs. Each query has a fixed random 256-example comparison bank excluding itself. Checkpoints at epochs 1, 5, 10, and 30 are evaluated at all five temperatures. This reuses images, but not corruption draws or trained heads; it is not independent-dataset replication. The prior experiment's results are development evidence for this protocol.

We evaluate the negative sample-to-mean-gradient cosine from AGRA's CE comparison rule at fixed checkpoints. Its adaptive update loop, alternative labels and F1 comparison loss are outside this experiment. The control averages negative pairwise cosines over the same noisy bank. This implements pairwise gradient comparison using noisy reference banks; the class-based detector of Nguyen-Duc et al. (2023) uses clean references. Label-only evidence uses 100 nearest neighbours. Self-gradient norm squared and its sum over recorded checkpoints are also evaluated as controls; neither is labelled a full TracIn reproduction. Confidence is stored as negative observed-label signed logit, preserving its binary temperature-invariant ranking without probability-saturation ties.

	AGRA-CE comparison AUROC at temperature T					Confidence
Epoch	0.25	0.5	1	2	4	AUROC
1	36.66	43.81	58.48	60.51	60.05	93.33
5	33.27	36.78	50.42	58.32	59.37	93.23
10	32.38	36.26	49.97	58.37	59.45	90.58
30	31.69	35.20	49.99	58.45	59.39	84.57

Table 14: All frozen checkpoints and temperatures, averaged over 15 runs. AUROC is reported in percent. Pairwise cosine remains 56.68% and label-only evidence 98.74% throughout; confidence rankings do not change with temperature.

The five final-checkpoint soft-minus-sharp AUROC gaps are $26.91, 33.77, 18.09, 30.83, 28.87$ pp. Pairwise cosine changes by at most 1.12×10^{-16}, with identical AUROC. At ordinary $T = 1$ checkpoints, The pairwise-cosine control improves over the AGRA-CE comparison in 48/60 cases, but loses to both confidence and label-only evidence in 60/60. Its mean differences from those two controls are -33.75 pp and -42.05 pp, respectively. Consequently the declared algorithm-promotion criterion is not met. The experiment confirms the mechanism prediction, while the control does not satisfy the stated criterion for further algorithm development.

M.7 Proof of polarity and pilot repair

Write the true class as $t \in \{-1, 1\}$, an observed label as y, and $v = (1, -1)$. Up to swapping class coordinates, $r = (\lambda t - y)v/2$. If $\overline{V}$ is the reference mean of ty_j, contextual work is $W = -(\lambda - ty)(\lambda - \overline{V})/2$. For independent noisy and clean AUROC samples,

$$2(W_{\text{noisy}} - W_{\text{clean}}) = (1 + \lambda)\overline{V}_{\text{noisy}} + (1 - \lambda)\overline{V}_{\text{clean}} - 2\lambda.$$

This proves Equation (31). The right side has mean $2(a - \lambda)$; its $2k$ independent summands have total squared ranges $8(1 + \lambda^2)/k$. Hoeffding's inequality proves both tail bounds. At $\lambda = a$, the nondegenerate centred bounded sum obeys a central limit theorem, giving limiting AUROC $1/2$. Merely observing that unscaled scores vanish would not prove the ranking conclusion.

For a pilot with aligned mean $\widehat{a}$, observable polarity is $\operatorname{sign}[\widehat{a}(\widehat{a} - \lambda)]$. Its desired value is $\operatorname{sign}(a - \lambda)$. An error requires either crossing λ or choosing a nonpositive orientation. Hoeffding bounds these events, including ties, by $e^{-md^2/2}$ and $e^{-ma^2/2}$. On correct pilot signs for both AUROC samples, the repaired comparison is the correctly oriented raw comparison. A union bound over the two pilots and the raw ranking error proves Corollary 5; negative lower bounds are interpreted as zero. In the multiclass homogeneous analogue, $\alpha = u + a(e_t - u)$ and $p = u + \lambda(e_t - u)$ give noisy-minus-clean population work gap $a - \lambda$, with $a = 1 - C\eta/(C - 1)$. The displayed finite-sample theorem is binary only.

M.8 Numerical checks and provenance

Both frozen tests save and SHA-256 hash their 15 score bundles before metric evaluation; evaluators verify input, protocol, code and score hashes against the manifests in Section R.3. For the first test, all 825 metric rows reproduce; independent scalar and vector implementations agree to 6.67×10^{-16}. Exact-binomial checks cover 27 configurations, multiclass population identities agree to 1.00×10^{-15}, and simulations preserve algebraic ties. For the cross-method test, an independent audit reconstructs all 2,100 AUROC and average-precision (AP) rows, with maximum score discrepancy 1.78×10^{-15} and metric discrepancy 3.34×10^{-16}. Analytic gradients match autograd below 4.45×10^{-16}. The criterion ledger was independently checked. These checks establish internal integrity; hashes do not supply independent timestamps.

M.9 Predeclared criteria and observed outcomes

This retrospective audit organises the recorded criteria from the earlier mechanism studies. Each unit is a stated criterion; seeds, temperatures, class pairs, and detector pairings are repetitions within it. We distinguish exact controls, empirical extensions of a mechanism, and proposed algorithmic benefits. The heterogeneous-image extensions below are motivated by the homogeneous theory; its assumptions do not guarantee their outcomes.

These criteria are complete within the two recorded protocols and share images, models and temperatures; they are not independent samples from a research-wide prediction registry. The promotion gate remains unsuccessful. The image-level extensions test the homogeneous theory empirically. Reference-composition, ordinary-checkpoint and cancellation audits are retrospective and retain that status.

N Supporting Proofs and Controlled Checks

N.1 Additive representation, converse, and obstruction

Take $g_s(u, z, c) = E_s(u, \{(z, c)\})$. Repeated multiset additivity gives Equation (14); singletons prove uniqueness. Conversely, splitting the sum over a multiset union proves additivity with multiplicities. If $A = KR$, then $Rv = 0$ implies $Av = 0$, proving necessity. For sufficiency define $K(Rv) = Av$ on $\operatorname{im} R$; kernel containment makes this well-defined and linear. Extend it to the finite-dimensional source space. Equivalently $K = AR^{\dagger}$, because $A = AR^{\dagger}R$. This construction applies to each query with the same R. All rows of all A must lie in the row space of R, proving the dimension lower bound; choosing a basis of their joint row span attains it. One-hot sources take $R = I_C$, $K = A$;

Criterion	Fixed prediction or decision rule	Recorded outcome
H1: extension	In at least four of five class-pair means, raw-work AUROC is above 55% at $T = 4$ and below 50% at $T = 0.25$.	Confirmed: five pairs.
H2: extension	At $T = 0.25$, orientation repair gains at least $5\,\mathrm{pp}$ in at least four pair means.	Confirmed: five pairs.
H3: benefit	At $T = 0.25$, the repaired combined score gains at least $0.1\,\mathrm{pp}$ over the 100-neighbour reference combined score, averaged over the fixed runs.	Confirmed: $3.001\,\mathrm{pp}$.
H4: control	Binary confidence AUROC is invariant across the fixed temperatures to 10^{-10}.	Confirmed: zero change.
P1: extension	At epoch 30, AGRA-CE comparison AUROC at $T = 4$ exceeds that at $T = 0.25$ by at least $10\,\mathrm{pp}$ in four pair means.	Confirmed: five pairs; gains 18.09–$33.77\,\mathrm{pp}$.
P2: control	Pairwise-cosine scores and AUROC are temperature-invariant to 10^{-10}; confidence and label AUROC are also invariant.	Confirmed: score change 1.11×10^{-16}; zero AUROC change.
Promotion gate	Promote the normalised-gradient control only if it improves beyond confidence and direct labels at ordinary checkpoints.	Not met: loses to each in all 60 comparisons.

Table 15: All named criteria H1–H4 and P1–P2 from the two polarity protocols, plus the second protocol's separate algorithm-promotion gate. The tolerances retain the original score units. The table includes the unsuccessful algorithm-promotion criterion alongside the confirmed mechanism tests.

$\|g_s(u, z, c)\| \le L$ gives $\|K_s(u, z)\|_{\mathrm{op}} \le \sqrt{C}L$. Finally $(A - KR)v = Av$ gives Equation (17). Identical aggregates force identical query outputs; the triangle inequality forces one of the two target errors to be at least $\delta/2$.

N.2 Trusted-label selection

Condition on the bank, score functions, and positive/negative evaluation counts. Replacing one positive observation changes empirical AUROC by at most $1/n_+$; replacing one negative changes it by at most $1/n_-$. The sum of squared bounded-difference constants is $1/n_+ + 1/n_-$. The empirical pair average is unbiased for population AUROC under the independent class conditional draws. McDiarmid's inequality and a union bound over L candidates give the displayed simultaneous bound. Comparing the selected candidate with the population maximiser uses the bound twice and gives 2ϵ regret; the ordering certificate follows in the same way.

N.3 Perfect detection

Conditioned on $U = u$, predicting one has error $1 - q^\star(u)$ and predicting zero has error $q^\star(u)$. Their minimum is $\min\{q^\star(u), 1 - q^\star(u)\}$. Taking expectation gives Equation (10). This expectation is zero exactly when its nonnegative integrand is zero almost surely, equivalent to $q^\star \in \{0, 1\}$ almost surely.

N.4 Representation loss

The Markov condition implies

$$q_Z = \mathbb{E}[D \mid Z, \widetilde{Y}] = \mathbb{E}[q_X \mid Z, \widetilde{Y}].$$

The function $v \mapsto \min\{v, 1 - v\}$ is concave on $[0, 1]$. Conditional Jensen followed by expectation therefore gives Equation (11).

For the second statement, conditional Pinsker gives

$$|q_X - q_Z| \le \sqrt{\frac{1}{2}\,\mathrm{KL}(\mathrm{Ber}(q_X) \,\|\, \mathrm{Ber}(q_Z))}.$$

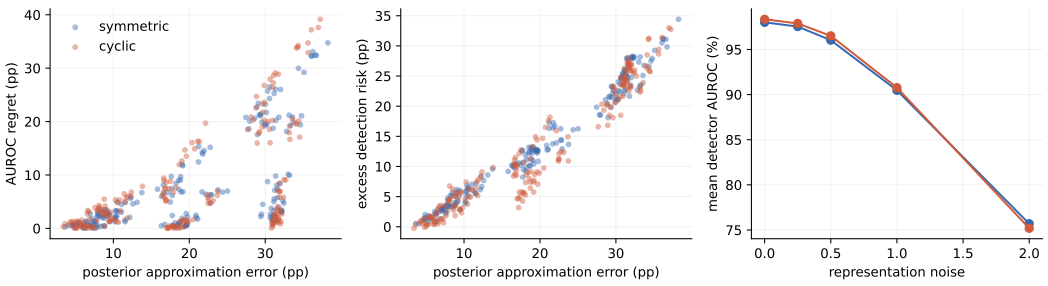


Figure 3: **Representation and query errors in the controlled study.** Left: total posterior error versus AUROC regret. Middle: the same error versus excess thresholded detection risk. Right: mean AUROC falls as representation noise increases, averaged over bandwidths and seeds.

Jensen's inequality for the square root and the definition of conditional mutual information yield

$$\mathbb{E}|q_X - q_Z| \le \sqrt{\frac{1}{2} I(D; X \mid Z, \widetilde{Y})}.$$

N.5 Plug-in detection risk

The two classifiers differ only when $q^\star$ and $\widehat{q}$ lie on opposite sides of $1/2$. The conditional excess risk is then $|2q^\star - 1|$. On the disagreement event,

$$|q^\star - 1/2| \le |q^\star - \widehat{q}|,$$

so the conditional excess is at most $2|q^\star - \widehat{q}|$. Integrating proves Equation (18).

N.6 AUROC regret

Condition on an unordered pair of feature values u, u'. If $q^\star(u) > q^\star(u')$, ordering u above u' rather than below it improves the probability of correctly ordering a positive–negative pair by $q^\star(u) - q^\star(u')$. Ties lose half this advantage. Averaging over the two orientations of independent U, U' and normalising by the positive–negative pair probability $2\pi(1-\pi)$ gives Equation (19).

If $s = \widehat{q}$ is uniformly within ϵ of $q^\star$, a reversed or tied pair must satisfy

$$|q^\star(u) - q^\star(u')| \le |q^\star(u) - s(u)| + |s(u') - q^\star(u')| \le 2\epsilon.$$

Restricting the expectation in Equation (19) to those pairs proves Equation (20).

O Controlled Decomposition and Neural-Identity Checks

O.1 Separating representation and query error

We draw four equiprobable clean classes from isotropic Gaussian components whose means lie on a circle. Labels are corrupted at rate 30% under either symmetric or cyclic class-conditional noise. Because the mixture and transition matrix are known, we evaluate the exact full-input posterior q_X from Equation (9). Representations are generated by adding Gaussian noise with standard deviation in $\{0, 0.25, 0.5, 1, 2\}$, for which q_Z is also analytic.

At each representation, a leave-one-out Gaussian-kernel query estimates the observed-label field. We vary its bandwidth over $\{0.2, 0.4, 0.8, 1.6, 3.2\}$ and use the known transition matrix to invert the observed field, project onto the simplex, and calculate $\widehat{q}$. The known transition matrix removes transition-estimation error, allowing this controlled experiment to isolate representation and query errors. Eight seeds, two noise mechanisms, five representations, and five bandwidths give 400 configurations.

The posterior decomposition and plug-in risk inequalities have no violations in the 400 rows. Total posterior error correlates with excess detection risk at $r = 0.972$ and with AUROC regret at $r = 0.639$

(Figure 3). Representation error is the dominant ranking diagnostic, correlating with AUROC regret at $r = 0.916$; query error alone has correlation 0.126. This difference is expected from Proposition 8: a biased or poorly calibrated posterior may preserve much of its ordering, whereas representation collapse can reverse large-gap pairs.

Mean AUROC falls from 98.2% to 75.4% as representation noise grows from 0 to 2, while query error remains approximately $13\,\mathrm{pp}$. Intermediate bandwidths perform best; a very broad kernel increases posterior error while preserving much of the ranking. These tests isolate the effects of representation and query error.

O.2 Exact neural interaction check

We instantiate a double-precision two-layer tanh network with three outputs at 20 random finite-width states. We form the full logit Jacobian and compare: (i) the autograd parameter gradient with the sum of example-local sources; and (ii) $-J\nabla_\theta \mathcal{L}$ with the queried field $-\Theta(p-y)/n$. Maximum relative errors are 4.11×10^{-16} and 5.10×10^{-16}, respectively. A step of size 10^{-4} has mean relative first-order remainder 6.66×10^{-6}. This checks the implementation of Theorem 8; the theorem itself is algebraic.

Together, the controlled experiments show how independently varied representation and query errors affect detection risk and ranking performance.

P Controlled Experiment Details

Clean field. Four class means are equally spaced on a circle of radius two. Classes have uniform prior and covariance $0.8^2 I_2$. The exact posterior is obtained by normalising the four Gaussian densities. A representation with noise scale σ_z adds independent $\mathcal{N}(0, \sigma_z^2 I_2)$ noise, so its exact class-conditional covariance is $(0.8^2 + \sigma_z^2) I_2$.

Corruption. Symmetric corruption retains a label with probability 70% and distributes probability 30% uniformly across the other three classes. Cyclic corruption retains it with probability 70% and maps it to the next class with probability 30%. Both transition matrices are constant and invertible. The experiment samples the latent and observed labels, but the detector receives only the representation and observed labels.

Query. For bandwidth h, the leave-one-out weights are

$$K_{ij} = \exp\left(-\frac{\|z_i - z_j\|^2}{2h^2}\right), \qquad K_{ii} = 0.$$

Their normalised weighted observed-label counts estimate $\mathbb{P}(\widetilde{Y} \mid Z = z_i)$. Multiplication by T^{-1} recovers a clean-field estimate, which is projected onto the probability simplex before applying Equation (9). AUROC uses average ranks for ties.

Reported errors. Representation error is $n^{-1} \sum_i |q_X(x_i, \widetilde{y}_i) - q_Z(z_i, \widetilde{y}_i)|$. Query error substitutes q_Z and $\widehat{q}$. Total error compares q_X with $\widehat{q}$. The decomposition bound is the sum of the first two errors. Excess risk compares posterior thresholding at $1/2$ with the full-input Bayes rule. Correlations are descriptive Pearson correlations across the full configuration grid; repeated configurations within seeds are not treated as independent hypothesis tests.

Neural identity check. For each of 20 seeds, we sample 24 six-dimensional inputs and a two-layer tanh network with hidden width 12 and three logits. The full logit Jacobian is constructed in double precision. We independently compute the mean cross-entropy gradient with autograd, its residual-source form $J^\top(p-y)/n$, the direct logit velocity $-J\nabla\mathcal{L}$, and the field query $-JJ^\top(p-y)/n$.

Q EXPLICIT DETECTOR WITNESSES

Each witness identifies the operation represented exactly and the state on which it conditions. Specifying these stages clarifies how their outputs become evidence for subsequent updates and readouts.

Neighbourhood label and residual evidence. Given a selected graph with weights w_{ij}, use sources $(e_{\widetilde{y}_j}, 1)$ and propagator $w_{ij}I$ to accumulate label counts and mass. The query divides by mass and compares class coordinates. For residual work, use sources $p_j - e_{\widetilde{y}_j}$; the query takes minus their aggregate's inner product with the query example's residual (Section L). One-hot and residual blocks instantiate Theorem 6 directly. Neighbourhood construction and out-of-fold prediction fitting are explicit preceding operations. Local consensus depends on representation geometry and correlated noise; higher-order consensuses can identify transitions under clusterability (Zhu et al., 2021; 2022).

Predictive scores and neural dynamics. Conditional on a fitted predictor, confidence, loss, and error ℓ_2-norm (EL2N) (Paul et al., 2021) are query-local functions of $p_i, \widetilde{y}_i$. Their dependence on other examples arises through training. Under the gradient flow assumptions in Theorem 8, the source is the class residual and the propagator is the tangent block. The self-term is explicit in Equation (27). For discrete SGD, the parameter update is exactly a sum of local gradients, while a finite logit update need not equal its first-order tangent expression. Training horizon and memorisation affect these predictive scores. AUM (Pleiss et al., 2020) and forgetting (Toneva et al., 2019) are exact readouts of the specified margin path under the conditions in Proposition 9; discrete checkpoint counts require their own checkpoint readout.

Confident learning: two aggregate passes. For confident learning (Northcutt et al., 2021), fixed out-of-sample probabilities give thresholds

$$t_c = \frac{\sum_j \mathbf{1}\{\widetilde{y}_j = c\} p_{j,c}}{\sum_j \mathbf{1}\{\widetilde{y}_j = c\}}.$$

Numerator and denominator are sums of local sources. With t fixed, let $a_j = \mathbf{1}\{\exists c : p_{j,c} \geq t_c\}$ and let $h_t(p_j)$ select the unique qualifying class, or the maximal-probability class when several qualify. Then the uncalibrated confident joint is

$$J = \sum_j a_j e_{\widetilde{y}_j} e_{h_t(p_j)}^\top.$$

It is an identity-propagated matrix source. Threshold tolerances, empty-class rules, row calibration, and issue selection are additional specified readouts. Calibration and transition misspecification affect the resulting estimates. Self-confidence is simply $p_{i,\widetilde{y}_i}$; it does not itself estimate this matrix. A global statistic can be expressed as its leave-one-out sum plus the query's local term. This gives concrete aggregate witnesses for the counting stages; the preceding predictor-fitting stage requires its own representation.

Influence and TracIn. Let $g_j = \nabla_\theta \ell_j$ at a fixed state and let H be a specified invertible Hessian. For a reference set with gradient mean $\bar{g}_R$, the usual influence approximation is $-\bar{g}_R^\top H^{-1} g_i$ (Koh & Liang, 2017). Sources are g_j, propagation is $H^{-1}/|R|$, and the query is $-g_i^\top F$ for symmetric H. This factorisation is exact for the computed influence approximation; equality to actual deletion and retraining effects is not asserted. TracIn (Pruthi et al., 2020) stacks checkpoint sources $(g_j^{(t)})_t$ and applies the block operator with blocks $\eta_t I$; the query contracts with $(g_i^{(t)})_t$. Self-influence retains the self-block. Curvature approximation and gradient scale affect these interactions. Reference-set choice, damping, checkpoint weights, and sign conventions belong to the supplied method.

Co-teaching: conditional update and selection feedback. For co-teaching (Han et al., 2018), let $S_B(t)$ be the examples selected at step t by model B for model A. With the mask frozen, the cross-entropy update for A is

$$\Delta\theta_A = -\frac{\eta_t}{|S_B(t)|} \sum_j \mathbf{1}\{j \in S_B(t)\} J_{A,j}^\top (p_{A,j} - e_{\widetilde{y}_j}).$$

The selected mask multiplies the propagator; SGD momentum can be carried as explicit state. Loss sorting constructs that mask and is not an additive operation. DivideMix (Li et al., 2020) adds mixture fitting, sample selection, and label construction; the taxonomy alone does not establish an exact representation of all these stages. This feedback can reinforce selection errors in both methods. Robust training without a corruption score is outside the detector class.

Fixed LEMoN. For LEMoN (Zhang et al., 2025), with image and text embeddings and neighbour graphs fixed, the equal-weight, zero-temperature specialisation scores direct image–label discordance plus two graph averages: observed-label disagreement over image neighbours and image disagreement over text neighbours. Label indicators, image embeddings, and a mass channel are sufficient sources; the query contracts with the query image or selects its label coordinate. Graphs depend on both modalities and must be specified as state. With a first-neighbour-skip convention, tied text embeddings can leave self-terms; these must be recorded explicitly as local information. The witness concerns the specified fixed specialisation; hyperparameter optimisation is a separate stage.

Coverage boundary. These witnesses establish conditional algebra for named components. The training-source characterisation in the main text extends this account to complete finite training paths in the specified fixed-feature setting. Further adaptive mechanisms require the same explicit accounting of how each stage changes the fitted state and its available evidence.

Q.1 Representation, mechanisms, and constructive consequences

George et al. (2024) organise detectors by base model, probe, ensemble, and aggregation, with executable component substitutions. Our components resolve a different level: the model and ensemble supply declared states; sources and propagation describe how examples contribute evidence within those states; queries and temporal readouts turn it into scores. These are complementary axes. A broad probe slot can describe either gradient comparison in Proposition 10. Our rescaling analysis predicts which comparison is invariant and when the other can reverse.

The algebra builds on established results. Kernel smoothing provides the normalised positive field (Nadaraya, 1964); graph label averaging shares the propagation primitive of Zhu & Ghahramani (2002), without identifying one step with their complete iterative algorithm. Zaheer et al. (2017) study sum-based invariant representations; Wagstaff et al. (2019) clarify regularity and dimension constraints. Our theorem assumes the stronger property of multiset additivity and asks whether a specified shared linear source preserves prescribed evidence. Its factorisation criterion uses standard kernel containment. The detector-specific contribution is the common characterisation, explicit stage boundaries, and consequences for source choice, decision preservation, and interaction mechanisms.

Calibration can already improve model-probing detection (Chibane et al., 2025). Our confidence intervention isolates a narrower distinction: positive binary-gradient rescalings cancel before aggregation when each source is normalised, but remain in an unnormalised aggregate and can reverse its direction. The intervention isolates score orientation at fixed classification decisions. The contribution connects these mechanism distinctions to exact information requirements, controlled intervention forecasts, and adaptive configuration search.

R Implementation and Experimental Specification

The repository's `README.md` documents data preparation and the training-search and polarity entry points; `SUPPORTING.md` gives supporting pipelines. Driver filenames below are relative to `experiments/`. The complete training-search settings appear in Sections G and H; this section specifies the shared image representation and polarity implementation.

R.1 Coordinates and prediction heads

The pretrained image representation is the 512-dimensional global-average pooling output of ResNet-18, ImageNet-1K weights `IMAGENET1K_V1` (He et al., 2016). Images are converted to RGB, resized with the pretrained transform (short side 256, bilinear interpolation), centre-cropped to 224×224, and normalised with channel means $(.485, .456, .406)$ and standard deviations $(.229, .224, .225)$. The

encoder remains frozen. The CIFAR pretrained features are cached in half precision and converted to single precision for queries. Dataset-specific raw-image preprocessing is held fixed within every paired comparison.

The polarity experiment's probability head is an affine d–128–64–C network with biases, rectified linear unit (ReLU) activations after each hidden layer, and softmax probabilities. It trains for ten epochs with cross-entropy, SGD learning rate 0.05, momentum 0.9, batch size 256, no weight decay, no dropout, and no learning-rate schedule. For seed s, a tensor-generator permutation with seed $s + 41000$ assigns alternating entries to the two held-out folds. The model for fold f is initialised and shuffled with seed $s + 100f$ and trained on the opposite fold. Every example receives only its held-out prediction.

R.2 Neighbourhoods and score evaluation

Nearest neighbours are obtained by L2-normalising feature rows and ranking their dot products. Set each self-similarity to negative infinity before selecting the largest k values. Search is exact and can be chunked without changing the mathematical query. The evaluated tensor implementation does not prescribe an index-based rule for equal feature similarities; average score ranks do not eliminate this possible neighbour-tie sensitivity. The label, work and combined scores are defined in Section L; the reference-bank split and temperature interventions are specified in Sections M.2 and M.6.

AUROC assigns half credit to ties, equivalently using the Mann–Whitney average-rank statistic. In the first polarity study, average precision is the mean precision at each corrupted example's position in descending score order. Exact ties are ordered by descending row index, as induced by reversing a stable ascending sort. Top-rate F1 uses the same order and inspects the known number of corruptions; it is an oracle-budget metric. The cross-method confirmation instead uses threshold-grouped average precision. These secondary metrics are not pooled across conventions; AUROC has the same tie convention throughout. Paired changes compare identical examples and corruption realisations. Shared images, checkpoints, temperatures and rankings remain dependent.

R.3 Numerical checks and provenance

The image implementation uses PyTorch 2.7.1 and torchvision 0.22.1. An NVIDIA A10 GPU instance was used for experimentation. Feature extraction and baseline training are separate from cached-score evaluation. Neighbour ties, floating-point arithmetic and hardware/library differences can affect bitwise reproduction. Independent calculations check source-factorisation identities, finite-network gradients, polarity algebra, score orientation and reported metrics.

Fresh runs create their own manifests. Under the default `results/polarity_prospective/` and `results/cross_method_polarity/`, `pre_score_manifest.json` records protocol, scoring-code and input hashes, and `score_manifest.json` additionally records all 15 score-bundle hashes. Evaluators verify these before computing metrics.

The supplied `REVIEW_ROBUSTNESS_PROTOCOL.md` fixes encoder seeds and cancellation filters. `review_robustness_audit.py` exports separate scores under `results/review_robustness/`; `validate_review_robustness.py` independently checks ranks and selections. Its input lock and score manifest retain SHA-256 identities.

Protocols and inputs are fixed before scoring; exported scores are hashed and verified before evaluation. This records an internal scoring/evaluation separation. Hashes establish content identity, without independent timestamps. Prospective claims concern the specified interventions and do not imply previously unused datasets.